Unsupervised domain adaptive multiple person tracking and visual identification for human-robot interaction

Guillaume Delorme

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 25 mai 2016

Présentée par

Guillaume DELORME

Thèse dirigée par **Radu HORAUD**, Directeur de recherche, Université Grenoble Alpes et
codirigée par **Xavier ALAMEDA-PINEDA,** Université Grenoble Alpes

préparée au sein du **Laboratoire Laboratoire Jean Kuntzmann** dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Adaptation de domaine non supervisée pour modèle de suivi multi-partie et identification visuelle appliquée à l'interaction homme-robot

Unsupervised domain adaptive multiple person tracking and visual identification for human-robot interaction

Thèse soutenue publiquement le **8 octobre 2021**,
devant le jury composé de :

Monsieur RADU HORAUD
DIRECTEUR DE RECHERCHE, INRIA, Directeur de thèse
Monsieur XAVIER ALAMEDA-PINEDA
CHARGE DE RECHERCHE, INRIA, Co-directeur de thèse
Monsieur ALBERTO DEL BIMBO
PROFESSEUR, UNIVERSITA DI FIRENZE, Rapporteur
Madame LOURDES AGAPITO
PROFESSEUR, UNIVERSITY COLLEGE LONDON , Rapporteure
Monsieur VINCENT LEPETIT
PROFESSEUR, ECOLE DES PONTS PARIS TECH, Président
Monsieur, STÉPHANE, LATHUILLIÈRE
MAÎTRE DE CONFÉRENCE, Examinateur Invité

Abstract

Human robot interaction requires the robot to have an accurate knowledge of its environment, especially who is present, and where, to enable an interactive conversation. In this context, this thesis proposes to exploit image information recorded by the embedded camera to perform Multiple Object Tracking (MOT), leveraging localization and identification by exploiting temporal and spatial proximity to produce ID-exploitable trajectories. State-of-the-art methods rely on deep learning approaches, which are known to heavily depend on the training data, and suffer from poor generalization ability. More specifically, most of MOT implementations embed a person re-identification model to use as appearance cue, while those are widely known to be sensitive to background changes and illumination conditions. Consequently, this work focuses on investigating adaptation strategies to new domains for MOT and re-ID models. A probabilistic generative model is first proposed to derive a MOT implementation which, combined with a deep appearance model updated with past track annotations, is able to adapt to the target domain on the fly, and is suitable for robotic application. It is quantitatively evaluated on a standard MOT dataset while a robotic implementation provides qualitative results. Then, inspired by the domain adaptation literature, a camera-wise adversarial strategy is proposed to address unsupervised person re-ID, and demonstrates competitive performance compared to state-of-the-art re-ID models. It is then further investigated in the novel framework of *clustering* and *finetuning*. A conditional adversarial approach is proposed to address the negative transfer problem caused by the non-uniform distribution of IDs across cameras. This strategy is implemented on two state-of-the-art unsupervised re-ID models, and shown to outperform them, thus yielding state-of-the-art performance. Finally, the adversarial domain adaptation framework is further investigated in the context of MOT. The interest for unsupervised domain adaptation MOT is demonstrated, and combined with a *tracking* and *finetuning* strategy, an adversarial training scheme is derived and shown to outperform simpler adaptation strategies.

Résumé

L'interaction homme-robot nécessite que ce dernier ait une connaissance précise de son environnement, tout particulièrement qui est présent et où, afin de permettre une conversation réaliste et interactive. À cette fin, cette thèse propose d'exploiter l'information contenue dans les images récoltées par la caméra du robot afin de réaliser du suivi multi-partie, utilisant la proximité temporelle et spatiale afin de produire des trajectoires exploitable à des fins d'identification. L'état de l'art est basé sur des approches d'apprentissage profond, qui sont connus pour dépendre grandement des données utilisées lors de l'entraînement, et ont donc une mauvaise capacité de généralisation à de nouveaux domaines. Une grande partie des modèles de suivis utilisent notamment des modèles de ré-identification de personnes comme descripteur d'apparence, alors que ceux-ci sont connu comme étant très sensible aux changements d'arrière plan, ou de conditions d'illumination. Ce travail se concentre donc sur l'investigation des stratégies d'adaptation à de nouveaux domaines pour les modèles de suivis et de ré-identification de personnes. Un modèle probabiliste est d'abord proposé pour implémenter un algorithme de suivi multi-partie qui, combiné avec un modèle d'apparance profond mis à jour en utilisant les annotations des trajectoires passées, est capable de s'adapter au domaine cible en temps réel, ceci dans un contexte robotique. Cette stratégie est quantitativement évaluée sur un dataset standard de suivi multi-partie, et une implémentation sur une plateforme robotique fournit des résultats qualitatifs. Ensuite, inspiré de la littérature de l'adaptation aux nouveaux domaines, une stratégie d'entraînement adversaire basée sur l'information de caméra est proposée dans le cadre de la ré-identification de personne non supervisée. Cette approche démontre une performance compétitive comparé à l'état de l'art en ré-identification. Cette approche est davantage explorée à travers la nouvelle stratégie de *partitionnement* et *d'entraînement*. Une variante conditionnelle est explorée pour atténuer le problème de transfert négatif, causé par la répartition non uniforme des identités d'entraînement sur les caméras. Cette idée est implémentée sur deux modèles de l'état de l'art, et permet de les améliorer. Enfin, le cadre de l'adaptation de domaine adversaire est explorée dans le contexte du suivi multi-partie, et combiné avec une stratégie de *suivi* et d'*entraînement*, un algorithme d'apprentissage est proposé, et sa supériorité vis à vis des stratégies d'adaptation concurrentes est démontrée.

ACKNOWLEDGMENT

I express my sincere gratitude to my supervisors, Dr. Radu Horaud and Dr. Xavier Alameda-Pineda, for their support during this thesis: they were always willing to discuss extensively the scientific problems I encountered, and taught me the thoroughness required by academic research. More importantly, they taught me how to formulate research questions and the methodology to answer them.

I would like to thank the members of my thesis committee for accepting to review this work, namely Pr. Alberto Del Bimbo, Pr. Lourdes Agapito, Pr. Vincent Lepetit and Dr. Stéphane Lathuilière. I really appreciated the questions asked during the defense, and the scientific discussion that arose from them. I would especially like to thank my examiners Pr. Alberto Del Bimbo and Pr. Lourdes Agapito for providing thorough and meaningful reviews improving this manuscript.

I am also grateful to the people I closely collaborated with: Stéphane Lathuilière for his never-ending curiosity and his willingness to explore subjects extensively, Xavier Alameda-Pineda for his mathematical thoroughness, abstraction ability and his persistent optimism. I would especially want to thank Yihong Xu with which I collaborated closely on numerous projects. His deep expertise in MOT, and ability to quickly understand and adapt to new problems lead to fruitful collaborations. More importantly, his dedication and motivation greatly contributed to these projects. I would also like to thank Yutong Ban, Guillaume Sarrazin and Luis Gomez Camara who were involved in the implementation aspects of this thesis. I would also like to thank Sylvain Guy and Benoît Massé for the informal scientific discussions we had, and even though we did not directly collaborate, those discussion helped me to grow as a researcher.

More generally, I would like to thank the staff at inria whose work directly or indirectly facilitated this thesis: especially Nathalie, Soraya and Jean-François, who greatly helped on the administrative and technical aspects of my research. Your help and availability were a great support during this work. I also include all the current and former staff of the perception/robotlearn team, namely Laurent, Simon, Pablo, Mostafa, Chris, Timothée, Matthieu, Alex, Zhiqi, Louis, Anand, Xiaoyu[2], Wen, Gaétan, Vincent and Dionyssos.

I would like to also thank my family and friends for helping me go through this journey: Arnaud, Sylvain, Arthur, Cyril, Maxime, Thomas, Yoann, Gaétan, Chems, Mehdi, Anaïs, and all the others I don't have the space to mention. I especially thank Arnaud for the mountains, Benoit for the boardgames and Yihong, Sylvain, Denys and Mathieu for the culinary discoveries. Special thanks to Daphnée for being there and supporting me all along this path.

CONTENTS

1 Introduction **11**

1.1 General context 11
1.2 Scientific context and motivation 12
1.3 Datasets 13
1.4 Contributions 15
1.5 Manuscript Structure 16

2 Online Appearance Learning for Robotic Person Tracking **17**

2.1 Introduction 17
2.2 Related Work 18
2.3 Online Multi-person Tracking 19
2.3.1 Egomotion-aware Multi-person Tracking Model 19
2.3.2 Variational Online Multi-person Tracking 20
2.4 Joint Tracking and Appearance Modeling 22
2.4.1 Unsupervised Deep Metric Learning 22
2.4.2 Probabilistic Appearance Model Update 23
2.5 Overall Tracking System 24
2.5.1 Deep Appearance Model Update 24
2.5.2 Birth and Visibility Processes 24
2.5.3 Robotic implementation with ROS 25
2.6 Experiments 27
2.6.1 Quantitative evaluation 27
2.6.2 Qualitative results 30
2.7 Conclusion 30

3 Camera Adversarial Unsupervised Re-ID **33**
3.1 Introduction . . . 33
3.2 Related Work . . . 35
3.2.1 Supervised person re-ID . . . 35
3.2.2 Unsupervised person re-ID . . . 35
3.2.3 Adversarial for Unsupervised re-ID . . . 35
3.3 Camera Adversarial Transfer (CAT) . . . 37
3.3.1 Domain confusion loss . . . 38
3.3.2 Label smoothing regularization for Outliers . . . 39
3.3.3 Overall training . . . 39
3.4 Experiments . . . 40
3.4.1 Protocol . . . 40
3.4.2 Comparison to State of the Art . . . 41
3.4.3 Ablation study . . . 42
3.5 Conclusion . . . 46

4 Conditional Adversarial Network For Unsupervised re-ID **51**
4.1 Introduction . . . 51
4.2 Related work . . . 52
4.2.1 Unsupervised person re-identification . . . 52
4.2.2 Negative transfer . . . 53
4.3 Clustering based Unsupervised Person Re-ID . . . 54
4.4 CANU-ReID: A Conditional Adversarial Network for Unsupervised Person re-ID . . . 55
4.5 Experimental Validation . . . 57
4.5.1 Evaluation Protocol . . . 57
4.5.2 Comparison with the State-of-the-Art . . . 60
4.5.3 Ablation Study . . . 61
4.5.4 Impact of CANU on Camera Information . . . 62
4.5.5 Impact of CANU on lost IDs . . . 64
4.5.6 Embedding visualisation . . . 66
4.6 Conclusion . . . 67

5 Domain Adaptive MOT **69**
5.1 Introduction . . . 69
5.2 Related Work . . . 70
5.2.1 Multiple Object Tracking . . . 70
5.2.2 Domain Adaptive Object Detection . . . 71
5.2.3 Unsupervised Person re-ID . . . 71
5.3 Methodology . . . 72
5.3.1 Notations and Tracker Overview . . . 72
5.3.2 Tracking and Finetuning . . . 73
5.3.3 Adversarial Sequence Alignment . . . 74
5.3.4 Adversarial learning for disentanglement . . . 75
5.3.5 Overall training . . . 76
5.4 Experimental results . . . 77
5.4.1 Implementation details . . . 77
5.4.2 Evaluation Protocol . . . 77
5.4.3 Results and discussion . . . 78
5.5 Conclusion . . . 80

6 Conclusion **81**
6.1 Summary . . . 81
6.2 Future research directions . . . 82
6.3 Publications and Submissions . . . 84

A Appendix: Adversarial learning for multi-domain classifier **85**

List of Figures **90**

List of Tables **92**

List of Algorithms **93**

CHAPTER 1

INTRODUCTION

1.1 GENERAL CONTEXT

Robotic systems have seen significant development in the last decade, motivated by hardware improvement, and industrial growing interests (Boston Dynamics, SoftBank Robotics or PAL robotics). Combined with the success of Deep Learning architectures [59, 90] leveraging the possibility to extract information from dense raw input data captured by cameras or microphones arrays, it paved the road for a more natural Human-Robot Interaction. Humanoid Robots are nowadays employed in a wide variety of places, such as hospitals, nursing homes, museums and even stores. However, their applications have very limited interactivity: they are often limited to providing information or directions, taking orders via a standard graphical interface, or running predefined routines to entertain patients or the audience. Progress needs to be done to increase the adaptability of the robot to conversational situations, and to the speakers.

This PhD takes place in the framework of multi-party conversations. Such conversations are particularly challenging due to a varying number of speakers, and to the complexity arising from the multiplication of possible addressees. The first and crucial task a robot has to perform to ensure a fully interactive conversation, is then to answer the following questions: 'Who is speaking?', 'To whom?', 'When?' and 'Where?'. Answering those questions gives the robot the ability to associate information to each specific speaker, know whose turn it is to speak, and thus detect when it is spoken to and when to speak. In addition to help the interactivity of the conversation, this information is a prerequisite for the robot to have a socially and culturally acceptable discussion. This thesis specifically aims to answer the questions 'Who?' and 'Where?': identifying and retaining the position of encountered potential speakers exploiting visual cues provided by the robot camera is a critical step for credible robot-interaction.

To answer those questions, Multiple Object Tracker (MOT) has been explored in order to exploit temporal and spatial proximity to transform noisy and anonymous detections into ID-exploitable trajectories. Appearance models based on person re-identification

(Re-ID) models can also be used to assist tracking, especially in the case of mutual or prolonged occlusions in order to reduce the risk of ID switches and track fragmentation. However, most recent development of MOT and re-ID models build upon Deep Learning frameworks, yielding impressive tracking and re-ID performance, but require a large amount of labeled data to generalize well, which often lead to poor performance when transferred to new domains. This is of critical importance when it comes to actually implementing such systems in real-life settings: potential users rarely have the resources to create custom datasets on which to finetune their models for every use-case. Furthermore, long-term tracking/re-ID can see its target domain shifts significantly from the original training domain: illumination conditions can change with time, weather or location, the robot can also move or be moved, leading to potential significant background and viewpoint changes. This motivates the central question of this thesis: *How can tracking and re-ID systems be adapted to new domains, without requiring a tedious and costly labeling step?*

A first strategy addresses MOT by using a probabilistic setting based on a generative model, i.e. not requiring training, and adapts its appearance model to the target environment on the fly (chapter 2). Another approach pretrains a model within the unsupervised domain adaptation framework (UDA), that is using labeled *source* data and unlabeled *target* data during training to optimize performance on the *target* domain. More specifically, this framework can be applied in the context of unsupervised person re-ID (chapters 3 and 4), or to domain adaptive MOT (chapter 5).

1.2 SCIENTIFIC CONTEXT AND MOTIVATION

This thesis was caried out at Inria Grenoble Rhône-Alpes in the Perception team, and supervised by Dr. Radu Horaud and Dr. Xavier Alameda-Pineda, and is part of the Vision and Hearing in Action (VHIA) project supported by the ERC Advanced Grant obtained by Dr. Radu Horaud. The aim of VHIA and the team research interest lies in developing methods to improve human-robot interaction in multi-party conversations. The typical target situation is the 'cocktail party' scenario, where people move and discuss freely in front of the robot, in an indoor environment. The robot is equipped with one or two (stereo) cameras and an array of microphones. Its goal is to recover high level information of the discussion, such as who is present, or whose turn it is to speak. It led the team to thoroughly investigate methods to extract high-level information from audio-visual cues such as audio-visual tracking, sound source localization and audio-visual speech enhancement.

In such scenarios, since the camera field of view is limited, the robot must often choose which person to look at and adapt its position accordingly. Doing so, it might lose sight on a person of interest from whom identity information must be retained in order to grasp the overall conversation. Also, the environment might be cluttered and people can easily disappear from the robot field of view. Therefore, one must develop robust techniques to retrieve and retain ID information using visual information in particular. In addition, the typical scenario can greatly vary in terms of illumination condition (time of the year,

day/night) or background (office, common room, corridors...), it must also be able to adapt easily to new and previously unseen domains, hence the focus of this thesis.

In recent years, Deep Learning attracted a lot of focus to extract information from raw and dense data, especially in computer vision, outperforming many traditional computer vision tasks such as object detection [38, 86], Classification[90] and Recognition tasks[49]. It particularly impacted the MOT community, leveraging deep appearance descriptors such as person re-ID models [71, 91] or detectors with higher detection performance [76]. The person re-ID community was also deeply impacted [119, 62], and despite the relative small training datasets to learn on, deep re-ID models showed surprisingly good performance, even surpassing human re-ID ability in some cases [116]. However, these models showed poor generalization ability when transferred to new domains [25, 28], thus limiting their interest in the context of MOT. In parallel, and due to deep learning hunger for labeled data, the field of unsupervised domain adaptation introduced new methods[34] to leverage unannotated datasets, but initially focused only on classification models. This thesis aims to close the gap between MOT, person re-ID and domain adaptation.

This thesis benefited from two robots in the team, namely NAO and LITO. They are equipped with cameras and microphone array. NAO is a humanoid robot platform developed by Aldebaran Robotics. It is equipped with a pair of stereo cameras, four microphones and several joint motors for moving head, arms and legs. LITO, developed by Samsung, is a robot-head equipped with one camera, a set of microphones, and two motors for moving the robot head. Both NAO and LITO are research prototypes manufactured to make progress on the research topics of the team.

1.3 DATASETS

Deep Learning frameworks are known to depend on the labeled data they train on: also training and testing datasets are extensively used in this thesis for experimental evaluation. Two kind of datasets are used: the first are MOT datasets, constituted by a few sequences recorded on a (possibly moving) camera and labeled with ID and bounding box positions at every frame; the second are person re-ID datasets, providing person detections, recorded over a dozen of fixed cameras, on a short period of time.

MOT Challenge Datasets Pedestrians tracking has been studied for years in the MOT community. The MOT Challenge datasets thus provides a standard benchmark for tracking pedestrians. It is comprised of the MOT17 [76] dataset and MOT20 [24] dataset, the former being constituted of 7 training sequences and 7 testing sequences, while the latter has 4 training and 4 testing sequences. MOT17 has 517 Ground Truth training tracks, and 759 testing tracks, while MOT20 has 2332 Ground Truth training tracks, and 1507 testing tracks. They involve various settings such as the static camera or moving camera, indoor scenario or outdoor scenarios. Noticeably, the MOT20 dataset is much closer to what could be expected from a camera surveillance video sequence, with very crowded

(a) MOT17 Dataset

(b) MOT20 Dataset

(c) Market-1501

(d) DukeMTMC-reID

(e) MSMT17

Figure 1.1: Datasets used in this thesis: (a) and (b) are MOT datasets, (c), (d) and (e) are re-ID datasets.

scenarios from a top view, while the MOT17 dataset is much closer to the floor level, and is close to what a robot navigating in the crowd could experience. Both datasets have ground truth trajectories provided along the training sequences, along with official public detections produced by standard person detectors (DPM [30], FRCNN [38], SDP [107]). Two tracking settings are then possible: *public detection* where no extra detection is used to generate tracks, or *private detection* where custom detections are provided to generate tracks. The MOT16 dataset is also used in this thesis, and refers to the same sequences as MOT17, but using detections provided by only the DPM detector.

Person re-ID datasets Market-1501 (Mkt) [120], DukeMTMC-reID (Duke) [89] and MSMT17 (MSMT) [101] are used to benchmark person re-ID models. In all three cases,

the dataset is divided into three parts: training, gallery, and query. The query and the gallery are never available during training and only used for testing. Mkt is composed of $1,501$ (half for training and half for testing) different identities, observed through 6 different cameras (viewpoints). The deformable parts model [30] is used for person detection. Duke is composed of $1,404$ (half for training and half for testing) identities captured from 8 cameras. In addition, 408 other ID, called "distractors", are added to the gallery. Detections are manually selected. MSMT is the largest and most competitive dataset available, with $4,101$ identities ($1,041$ for training, and $3,060$ for test) on 15 cameras, with a much higher temporal coverage and background changes than previous ones .

1.4 CONTRIBUTIONS

The contribution this thesis are the following:

- The problem of MOT adaptability in robotic settings is first investigated: a probabilistic framework is leveraged to derive a generative model, which does not require pretraining and thus adapts easily to new target domains. It can be seen as a generalization of Kalman filtering to MOT, combined with a variational approximation to make the model tractable. It continuously updates an appearance model to the target experiment, by using a deep siamese network trained with a constrative loss and self-supervised with past detections annotated by the tracker. This MOT strategy is implemented on a real robotic platform to provide qualitative tracking results, and is further validated on MOT16 and MOT17 to obtain quantitative results. Importantly, the robotic implementation runs in a real-time setting, reaching an average of 10FPS.
- The problem of Unsupervised Person re-ID is investigated in the framework of Unsupervised Domain Adaptation: the *source* labeled Re-ID dataset and *target* unlabeled Re-ID dataset are both used during training to improve re-ID performance on the *target* domain. A camera-adversarial strategy is derived to perform camera-wise distribution alignment instead of domain-wise distribution alignment. The adaptation method is evaluated on standard unsupervised re-ID settings, Market1501 $\rightarrow$ DukeMTMC and DukeMTMC $\rightarrow$ Market1501, demonstrating competitive performance compared to state-of-the-art methods at the time, and an extensive ablation study is performed to compare with alternative adversarial strategies.
- The problem of Unsupervised Person re-ID is then investigated within the *Clustering* and *Finetuning* framework, a setting closer to standard unsupervised learning. A camera-adversarial strategy is leveraged to guide the clustering step while ensuring camera-wise distribution alignment. The impact of negative transfer is investigated, that is the impact of non-uniform ID distribution across cameras when using a camera-adversarial strategy. A conditional adversarial strategy is developed to address negative transfer. The effectiveness of the method is demonstrated by improving two state-of-the-art *Clustering* and *Finetuning* re-ID models on Market1501, DukeMTMC and MSMT re-ID datasets, and an extensive ablation study is provided to assess the role of conditional adversarial learning.

- Domain Adaptive MOT is investigated to tackle domain-shift problems when transferring Deep MOT models to new illuminations conditions, background or camera viewpoint. A Deep Tracker, jointly performing detection and re-ID within a unified model, is adapted using a proposed *tracking* and *finetuning* framework. The domain adaptive object detection framework is leveraged to finetune the detection branch. The interest of Domain Adaptive MOT is demonstrated on the MOT17 and MOT20 datasets, and the effectiveness of the proposed architecture is demonstrated on the MOT17→MOT20 and MOT20→MOT17 settings, while extensive ablation study is conducted to assess the role of each adaptation module.

1.5 MANUSCRIPT STRUCTURE

The manuscript is organized as follows. In chapter 2 the MOT robotic tracking system is described. In chapter 3 a camera-adversarial strategy is proposed to address the problem of unsupervised person re-ID. In chapter 4, the previous idea is further explored within the *clustering* and *finetuning* framework, and a conditional approach is proposed to address negative transfer. In chapter 5, a domain adaptive strategy for MOT is explored. In appendix A, a theoretical derivation is provided to justify the use of multi-class discriminator in adversarial learning.

CHAPTER 2

ONLINE APPEARANCE LEARNING FOR ROBOTIC PERSON TRACKING

2.1 INTRODUCTION

Multiple object tracking (MOT) is a well studied problem in the computer vision community [14, 76, 4, 91, 5], and the task is relevant for many applications, going from active surveillance to social robotics as detailed in Chapter 1. Several paradigms can be used to tackle that problem, and a large body of the literature is based on the tracking-by-detection philosophy, consisting on the use of a person bounding box detector [107, 30, 87], followed by a tracking model aggregating the detections over time, thus producing tracks. Such tracking models can be based on deep learning approaches [91, 21] or Bayesian [4, 14] models. This latter approach leverages generative models to tackle MOT, thus reducing the domain-dependency problem that learned discriminative frameworks might encounter, as developed in Chapter 5.

In this chapter, we are interested in endowing a domain-agnostic autonomous system with the ability of tracking multiple persons in an online setting, and in particular to update its appearance description model to the persons in the scene. This is challenging because of four main reasons: (i) the method can only use causal information, since the system does not have access to future bounding boxes; (ii) the method must be computationally light, in the sense that the system must track people using consumer technology; (iii) the model update must be done online in an unsupervised manner, since the system does not have access to ground truth annotations of the tracked people; and (iv) the overall system must account for visual clutter e.g. visual occlusions and image noise due to the system's ego-motion.

To that aim, we propose a probabilistic model combined with a deep appearance model. While the probabilistic model sets the relationship between the latent variables (e.g. people's position) and the observations (bounding boxes), the appearance model based on a deep siamese neural network allows to robustly discriminate images belonging to different people. Most importantly, and this is the main contribution of the chapter,

the probabilistic-deep siamese combination allows to update the deep appearance model with the supervision generated by the probabilistic model, avoiding the necessity of annotated data. This combination is the key that allows the update of supervised discriminative models in unsupervised settings, such as the task at hand.

Up to our knowledge, we are the first to propose a MOT method able to update a deep appearance architecture demonstrating its capabilities running in an autonomous robot. Indeed, we benchmarked our method with the state-of-the-art on standard datasets and standard evaluation procedure, and under two different settings: moving surveillance camera and robot navigation in crowded scenes. In addition, we provide qualitative results obtained on a real robotic platfom. The reported experiments validate our initial thoughts and confirm that updating the appearance model with the supervision from the probabilistic model is a good strategy. Indeed, the proposed method exhibits a significant performance increase when compared to the use of a fixed deep appearance model, and to the state-of-the-art. Our strategy appears to be effective for learning a discriminative appearance model on the fly, while tracking multiple people and accounting for clutter at the same time.

The Chapter is structured as follow: Section 2.2 does a review of the related work, Section 2.3 introduces the notations and the tracker's probabilistic model, Section 2.4 details the appearance model and how it is updated, and Section 2.5 discusses the practical implementation of the tracker on a robotic platform. Finally, Section 2.6 details the experimental evaluation of the algorithm.

2.2 Related Work

Tracking by detection is the most popular paradigm in the MOT community. Causal tracking is generally performed by elaborating robust similarity measures between known tracks and current detections, and by using data association methods to obtain optimal track-to-detection assignments. The major differences rely in the similarity measure used, which strongly depends on the visual cues that are used (e.g. spatio-temporal, appearance, interaction models).

Spatio-temporal similarity generaly assumes linear motion model [4, 14, 9]. The recent introduction of deep learning has leveraged the use of reliable appearance descriptors, allowing causal tracking models to robustly evaluate appearance similarities [60, 5]. The introduction of person Re-ID models [10, 71, 91] has allowed tracking algorithms to take advantage of external dataset to improve their generalization ability. Several strategies are developped to aggregate those similarity measures: [4] takes advantage of the probabilistic formulation to merge different cue information, while some [91, 21] propose a deep formulation where cues are merged early in the network which is directly trained to output the desired similarity measure. Single object tracking strategies [21, 5, 31] have also been exploited to perform multiple object tracking, especially [21, 5] using the siamese formulation which allows for an online finetuning of the appearance model to the target visual domain.

While the progress in MOT is quite significant, methods able to perform online MOT on autonomous robots are much more scarce. Computational complexity and moving cameras are two of the main difficultes most methods have trouble overcoming. One example of a MOT method fully adapted to robotic platforms is [7], where the authors propose a probabilistic model and a variational approximation to solve the tracking problem, while using the motor position to improve the tracking results. However, the appearance model used in [7] is based on color histogram descriptors, which lack or robustness and description power, specially in challenging visual conditions and for unseen identities.

We propose to exploit the same tracking formulation, to provide supervision for training a deep appearance models. Indeed, we exploit a Siamese deep network and use a soft-label formulation within the deep metric learning paradigm, to update the deep appearance model while tracking multiple people.

2.3 Online Multi-person Tracking

The proposed multi-person joint tracking and appearance model and its online robotic implementation, are built on the probabilistic tracking framework described in [4, 7]. In this section, we shortly describe the tracking model, so as to set up the discussion of the appearance model in Section 2.4.

2.3.1 Egomotion-aware Multi-person Tracking Model

In this chapter, let N denotes the maximum number of people to be tracked. $n = 0$ represents the clutter track.

State dynamics We want to infer the kinematic state for each person: $\mathbf{X}_{tn} = (\mathbf{L}_{tn}^\top, \mathbf{U}_{tn}^\top)^\top$, where $\mathbf{L}_{tn} \in \mathbb{R}^4$ is the person's bounding box (2D position, width and height) and $\mathbf{U}_{tn} \in \mathbb{R}^2$ is the person's velocity. $\mathbf{X}_t \in \mathbb{R}^{6N}$ is the person-wise concatenation of $\mathbf{X}_{tn}$. The dynamic model writes: $p(\mathbf{X}_t|\mathbf{X}_{t-1}) = \prod_{n=1}^{N} p(\mathbf{X}_{tn}|\mathbf{X}_{t-1n})$ where each person follows a Gaussian distributions:

$$p(\mathbf{X}_{tn}|\mathbf{X}_{t-1n}) = \mathcal{N}(\mathbf{X}_{tn};\ \mathbf{D}\mathbf{X}_{t-1,n} - \mathbf{E}_t, \mathbf{\Lambda}_n)$$

being $\mathbf{D}$ the dynamics operator adding the velocity to the previous position, and $\mathbf{\Lambda}_n$ accounting for dynamics noise. In order to compensate for the ego-motion of the autonomous system, we denote by $\mathbf{E}_t$, the impact of the motion of the system on the image plane at time t. See Section 2.6 for details on estimating $\mathbf{E}_t$.

Observation likelihood We assume that K_t bounding boxes are detected at time t. $\mathbf{y}_{tk} \in \mathbb{R}^4$ denotes the coordinates (2D position, width, height) of the k-th bounding boxes at time t, and $\mathbf{h}_{tk} \in \mathbb{R}^{w_k \times h_k \times 3}$ denotes its image content. We also write $\mathbf{o}_{tk} = \{\mathbf{y}_{tk}, \mathbf{h}_{tk}\}$ and $\mathcal{O}_t = \{\mathbf{o}_{t1} \dots \mathbf{o}_{tk} \dots \mathbf{o}_{tK}\}$. We assume that each bounding box is either generated

by a person or clutter, and denote by $Z_{tk} \in \mathcal{Z}_t \subset \{0, \dots, N\}$ the assignment variable associated to $\mathbf{o}_{tk}$. The likelihood writes:

$$p(\mathcal{O}_t|\mathcal{Z}_t, \mathbf{X}_t) = \prod_{k=1}^{K_t} p(\mathbf{y}_{tk}|Z_{tk}, \mathbf{X}_t) p(\mathbf{h}_{tk}|Z_{tk}).$$

The geometric likelihood is expressed as:

$$p(\mathbf{y}_{tk}|Z_{tk} = n, \mathbf{X}_t) = \mathcal{U}(\mathbf{y}_{tk})^{\delta_{0n}} \mathcal{N}(\mathbf{y}_{tk};\ \mathbf{P}\mathbf{X}_{tn}, \mathbf{\Sigma})^{1-\delta_{0n}},$$

meaning that if an observation is generated from a person, it follows a Gaussian distribution centered on the projection of the kinematic state using $\mathbf{P} = [\mathbf{I}_4, \mathbf{0}_{4\times 2}]$, and otherwise it follows an uniform distribution noted $\mathcal{U}$.

The appearance likelihood operates on $\phi_{\boldsymbol{w}}(\mathbf{h}_{tk}) \in \mathbb{R}^A$:

$$\begin{aligned} p(\mathbf{h}_{tk}|Z_{tk} = n) =& (\mathcal{U}(\phi_{\boldsymbol{w}}(\mathbf{h}_{tk}))^{\delta_{0n}} \times \\ & (\mathcal{N}(\phi_{\boldsymbol{w}}(\mathbf{h}_{tk}); \boldsymbol{m}_n, s_n^2 \mathbf{I}_A)^{1-\delta_{0n}}, \end{aligned} \tag{2.1}$$

where $\phi_{\boldsymbol{w}}(\cdot)$ denotes a feature extractor with parameters $\boldsymbol{w}$ and output dimension A, e.g. a CNN. Moreover, $\boldsymbol{m}_n \in \mathbb{R}^A$ represents a mean feature vector, and $s_n > 0$ is the concentration factor of the co-variance matrix. Both parameters need to be further estimated.

Finally, the prior distribution of the assignment variables decomposes as $p(\mathcal{Z}_t) = \prod_{k=1}^{K_t} p(Z_{tk})$ with:

$$p(Z_{tk} = n) = \rho_{tkn} \quad \text{and} \quad \sum_{n=0}^{N} \rho_{tkn} = 1, \tag{2.2}$$

where ρ_{tkn} represents the prior probability of $Z_{tk} = n$, and is a parameter that needs to be estimated.

2.3.2 VARIATIONAL ONLINE MULTI-PERSON TRACKING

The goal is to track multiple persons exploiting a probabilistic formulation. Since we pursue an online tracking algorithm, the model is constrained to use causal observations:

$$p(\mathbf{X}_t|\mathcal{O}_{1:t}) = \sum_{\mathcal{Z}_t} p(\mathcal{Z}_t, \mathbf{X}_t|\mathcal{O}_{1:t}), \tag{2.3}$$

$$p(\mathcal{Z}_t, \mathbf{X}_t|\mathcal{O}_{1:t}) \propto p(\mathcal{O}_t|\mathcal{Z}_t, \mathbf{X}_t) p(\mathcal{Z}_t)\ p(\mathbf{X}_t|\mathcal{O}_{1:t-1}). \tag{2.4}$$

The equation (2.4) is called *filtering distribution*, and all of the terms are already defined except for $p(\mathbf{X}_t|\mathcal{O}_{1:t-1})$, known as the *predictive distribution*.

The exact form of the filtering distribution leads to a computationally prohibitive strategy, and inspired by [7] we opt for a variational approximation:

$$p(\mathcal{Z}_t, \mathbf{X}_t|\mathcal{O}_{1:t}, \mathbf{\Theta}^\circ) \approx q(\mathcal{Z}_t, \mathbf{X}_t) = q(Z_t) q(\mathbf{X}_t), \tag{2.5}$$

which operates on the filtering distribution. It can be proven, that if the predictive distribution, the last term of (2.4), is a person-separable Gaussian distribution, then the predictive distribution at next time step will also be.

The optimal values for the variational distribution $q(Z_{tk} = n)$ and $q(\mathbf{X}_t)$ are obtained by minimizing the Kullback-Leibler divergence with respect to the exact filtering distribution. For the assignment variable we obtain:

$$\alpha_{tkn} = q(Z_{tk} = n) = \frac{\eta_{tkn}\rho_{tn}}{\sum_m \eta_{tkm}\rho_{tm}}, \tag{2.6}$$

where

$$\eta_{tkn} = \begin{cases} \mathcal{U}(\mathbf{y}_{tk})\mathcal{U}(\mathbf{h}_{tk}) & \text{if } n = 0 \\ \mathcal{N}(\mathbf{y}_{tk}, \mathbf{P}\boldsymbol{\mu}_{tn}, \boldsymbol{\Sigma})\mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\mathbf{P}^\top \boldsymbol{\Sigma}^{-1}\mathbf{P}\boldsymbol{\Gamma}_{tn})}\mathcal{N}(\phi_{\boldsymbol{w}}(\mathbf{h}_{tk}), \boldsymbol{m}_n, s_n\mathbf{I}_H) & \text{otherwise} \end{cases} \tag{2.7}$$

$\boldsymbol{\mu}_{tn}$ and $\boldsymbol{\Gamma}_{tn}$ are the parameters of $q(\mathbf{X}_t)$, which turns out to be a Gaussian distribution, and are recursively computed using:

$$\boldsymbol{\Gamma}_{tn} = \left(\tilde{\boldsymbol{\Gamma}}_{t-1n}^{-1} + \mathbf{P}^\top \boldsymbol{\Sigma}^{-1}\mathbf{P}\sum_{k=1}^{K_t}\alpha_{tkn}\right)^{-1}, \tag{2.8}$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn}\left(\tilde{\boldsymbol{\Gamma}}_{t-1n}^{-1}\mathbf{D}\boldsymbol{\mu}_{t-1n} + \mathbf{P}^\top \boldsymbol{\Sigma}^{-1}\sum_{k=1}^{K_t}\alpha_{tkn}\mathbf{y}_{tk}\right), \tag{2.9}$$

where $\tilde{\boldsymbol{\Gamma}}_{t-1n} = \mathbf{D}\boldsymbol{\Gamma}_{t-1n}\mathbf{D}^\top + \boldsymbol{\Lambda}_n$.

While (2.6) reminds of the E-step of an EM for GMM, (2.8) and (2.9) are the equivalent of the recurrent update of N parallel Kalman filters. The parameters of the model can be updated, as described in [7].

2.4 JOINT TRACKING AND APPEARANCE MODELING

In the previous section, we introduced the use of a non-linear mapping $\phi_{\boldsymbol{w}}(\cdot)$, which provides an embedding of appearance information using a CNN. Although CNN features for describing person appearances are already available, we expect that a higher representation performance will be exhibited if an already trained CNN is being fine-tuned using the bounding boxes and associated appearance information of those persons being actually tracked. In this section we describe a framework for coupling CNN fine-tuning with the proposed online tracker.

2.4.1 UNSUPERVISED DEEP METRIC LEARNING

Metric learning consists in learning a distance such that similar elements are close, and dissimilar ones are far apart. Siamese networks became a common framework for metric learning, also in the tracking community [48, 12, 60]. However, how to jointly update the parameters of the tracker and of the deep neural network is still an open question [60]. We first introduce the basics of Siamese networks in the context of tracking, to further on discuss the joint optimization.

Differently from classification problems, training Siamese networks requires a data set of triplets $(\mathbf{h}_i, \mathbf{h}_j, c_{ij})$, where $i, j \in \{1, \ldots, I\}$. The two bounding box images are feed-forwarded with the same weights $\boldsymbol{w}$, thus obtaining $\phi_{\boldsymbol{w}}(\mathbf{h}_i)$ and $\phi_{\boldsymbol{w}}(\mathbf{h}_j)$ respectively. The label is $c_{ij} = 1$ if the two images $\mathbf{h}_i$, $\mathbf{h}_j$ belong to the same person, and $c_{ij} = -1$ otherwise. A popular loss for training Siamese networks is a variant of the contrastive loss [41], introduced in [47]:

$$J(\boldsymbol{w}) = \frac{1}{2} \sum_{i,j=1}^{I} g(c_{ij}(\tau - \|\phi_{\boldsymbol{w}}(\mathbf{h}_i) - \phi_{\boldsymbol{w}}(\mathbf{h}_j)\|^2)), \tag{2.10}$$

with $g(x) = \max(0, 1 - x)$ and $\tau > 0$ is a parameter.

Traditionally, (2.10) is optimized with stochastic gradient descent, thus forcing the squared distance between elements from a negative pair to be higher than $\tau + 1$, and between a positive pair lower than $\tau - 1$. At test time, one can use the distance between the embedding vectors to gauge whether two appearances belong to the same person or not.

The metric learning framework has proven to be useful for a variety of tasks, but requires an annotated dataset. Our original motivation was to fine-tune the appearance model with the appearances of the people being tracked. By definition, such annotated dataset does not exist in our scenario. Therefore we need to adapt the deep metric learning methodology to our unsupervised setting.

In order to do that, we resource the deep metric learning problem with the information extracted by our online multi-person Bayesian tracking formulation. Indeed, (2.6) provides the posterior probability of the observation-to-person assignments: $\alpha_{tkn} = q(Z_{tk} =$

n) embodies the probability of observation k to be generated from the n-th target at time t. For any given two past observations k_i and k_j, from time frames t_i and t_j, we can compute the probability to be both generated by the same person:

$$\begin{aligned}\gamma_{ij} &= p(Z_{t_i k_i} = Z_{t_j k_j}) = \sum_{n=1}^{N} p(Z_{t_i k_i} = Z_{t_j k_j} = n) \\ &= \sum_{n=1}^{N} q(Z_{t_i k_i} = n) q(Z_{t_j k_j} = n) = \sum_{n=1}^{N} \alpha_{t_i k_i n} \alpha_{t_j k_j n}.\end{aligned} \tag{2.11}$$

Once the γ_{ij} are computed, we sample pseudo-positive pairs from those with $\gamma_{ij} > \eta$ and pseudo-negative pairs from those with $\gamma_{ij} < 1 - \eta$, respectively $\mathcal{H}^+$ and $\mathcal{H}^-$ sets. We optimize the following soft-weighted contrastive loss:

$$\begin{aligned}J(\boldsymbol{w}) = &\frac{1}{2} \sum_{ij \in \mathcal{H}^+} g(\gamma_{ij}(\tau - \|\phi_{\boldsymbol{w}}(\mathbf{h}_i) - \phi_{\boldsymbol{w}}(\mathbf{h}_j)\|^2)) \\ &- \frac{1}{2} \sum_{ij \in \mathcal{H}^-} g((1 - \gamma_{ij})(\tau - \|\phi_{\boldsymbol{w}}(\mathbf{h}_i) - \phi_{\boldsymbol{w}}(\mathbf{h}_j)\|^2)).\end{aligned} \tag{2.12}$$

2.4.2 Probabilistic Appearance Model Update

Simultaneously to the training of the Siamese network, one needs to update the appearance parameters of the probabilistic model, i.e. $\boldsymbol{m}_n$ and s_n^2, as defined in (2.1). As is the case for the parameters of the Siamese network, there is no annotated dataset, and this update must be done in an unsupervised manner. However, we still have access to the posterior probability of the observation-to-track assignments, and easily obtain the following updates:

$$\begin{aligned}\boldsymbol{m}_n &= \sum_{t'=t-w}^{t} \sum_{k=1}^{K_{t'}} \alpha_{knt'} \phi_{\boldsymbol{w}}(\mathbf{h}_{t'k}) \Big/ \sum_{t'=t-w}^{t} \sum_{k=1}^{K_{t'}} \alpha_{knt'} \\ s_n^2 &= \sum_{t'=t-w}^{t} \sum_{k=1}^{K_{t'}} \alpha_{knt'} \|\phi_{\boldsymbol{w}}(\mathbf{h}_{t'k}) - \boldsymbol{m}_n\|^2 \Big/ A \sum_{t'=t-w}^{t} \sum_{k=1}^{K_{t'}} \alpha_{knt'}\end{aligned} \tag{2.13}$$

with w being a moving window parameter. Updating these parameters at each time step allows the appearance model to be more flexible to sudden appearance variations and to better adapt the internal track appearances to the observations.

2.5 OVERALL TRACKING SYSTEM

In this section we describe the implementation of the method for joint tracking and appearance update, within a generic robotic platform operating under ROS.

2.5.1 DEEP APPEARANCE MODEL UPDATE

We instantiate our appearance model with a generic CNN backbone, consisting on several convolutional layers (see details below). In order to minimize (2.12), we use stochastic gradient descent with RMS [36] with a learning rate of $\lambda = 0.001$. The use of the entire training set $\mathcal{H}^+$ and $\mathcal{H}^-$ is prohibitive in terms of memory, and mini-batch SGD must be used. To construct batches, we randomly sample a track, one positive pair (from $\mathcal{H}^+$) and two negative pairs (from $\mathcal{H}^-$) associated to this track. This is done so as to respect the positive-negative balance [5, 21]. This strategy is repeated B times, obtaining a batch annotated with γ_{ij}.

$\phi_{\boldsymbol{w}}$ is pretrained to perform an ID classification similarly to [71]. Thus, the appearance model update can now be seen as a domain adaptation problem, where we need to learn the appearance shift (different people, background and illumination changes) between the pre-training dataset and the tracking data. To achieve this adaptation, only the top layers of $\phi_{\boldsymbol{w}}$ are updated during tracking. The amount of layers to be trained depends on the computation power of the system, allowing the best trade-off between generalization ability and computation complexity. In our case, we perform an update of the last 2 layers. We split the update into two steps. First, the *feature extraction* corresponds to the feed-forward through the frozen layers of $\phi_{\boldsymbol{w}}$. Second, the update of the trainable layers from the extracted features and γ_{ij}.

2.5.2 BIRTH AND VISIBILITY PROCESSES

New tracks (e.g. people coming in the field of view) are initialized using a birth process that testes for consistency the observations previously assigned to the clutter virtual track. We compare two hypothesis: (i) the previous L geometric observations $\mathbf{y}_{tk_0}, \ldots, \mathbf{y}_{t-Lk_L}$ assigned to clutter correspond to an undetected track, and (ii) the very same observations belong to clutter and are uniformly distributed. If the first hypothesis wins, a new track is initialized using detected bounding box, and its appearance model is initialized with the content of the bounding boxes using (2.13).

We used an additional hidden Markov model visibility process to determine whether or not a track has been lost. The observation of this binary visibility process arises from the output of the variational EM algorithm, in particular we set: $\nu_{tn} = \sum_k \rho_{tkn}$, representing whether a given track n is assigned to a current detection or not. The estimation of the latent variable probability $p(V_{tn}|\nu_{1:t})$ is done using standard HMM inference algorithms, and informs us about the visibility state of the considered track.

while *tracking* **do**
  $t \leftarrow t + 1$;
  Update $\{\alpha_{tkn}\}_{k\in[1,K_t],n\in[1,N]}$ with (2.6);
  Update $\{\boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}\}_{n\in[1,N]}$ with (2.8) and (2.9);
  Update $\{\boldsymbol{m}_n, s_n\}_{n\in[1,N]}$ with (2.13);
  if $t_u = T$ **then**
    $t_u \leftarrow 0$ // Reset frame counter;
    while $\phi_{\boldsymbol{w}}$ *not converged* **do**
      $\gamma_{ij} \leftarrow$ compute with (2.11);
      $\mathcal{H}^+, \mathcal{H}^- \leftarrow$ sample pair sets, section 2.5.1;
      Optimise (2.12) with RMS;
    end
  end
  $t_u \leftarrow t_u + 1$;
  Birth/visibility update, section 2.5.2;
  Output $\{\boldsymbol{\mu}_{tn}\}_{n\in[1,N]}$;
end

Algorithm 1: Overall tracking algorithm. Updates are performed with the various equations and strategies already described. The frame update counter t_u allows to update the $\phi_{\boldsymbol{w}}$ every T frames. The algorithm outputs the position of all tracks at every frame t.

2.5.3 Robotic implementation with ROS

The overall tracking algorithm is presented in Algorithm 1. While tracking, the algorithm uses the various updates derived from the variational EM algorithm (see Section 2.3). Every T frames, the system updates the appearance model with the equation updates of Section 2.4 and the sampling strategies and implementation details of Section 2.5.1. The birth and visibility processes of Section 2.5.2 are then used to set up new tracks and freeze non-visible tracks.

Algorithm 1 is implemented on a moving robotic platform using the ROS middleware. ROS does not only allow a platform-independent implementation (feature that will be used in our benchmark presented in Section 2.6), but also provides a unified framework to distribute the computation when and where needed. We use this property to exploit the computational power of an external GPU, devoted to execute the face detector [1] and to extract CNN appearance features. All other computations, including the update of the Siamese network, are ran on the CPU of the robot. We use a Intel(R) Xeon(R) CPU E5-2609 and a NVIDIA GeForce GTX 1070, and exploit the native camera of the robot. The system runs under Ubuntu 16.04 and ROS Kinetic version. Thanks to these implementation choices, our online tracker runs at 10 FPS. A schematic representation of the overall tracking system is shown in Figure 2.1.

ROS makes use of a distributed network of Nodes (ROS processes), which use topics to communicate. Typically, the drivers nodes control the low level communication with

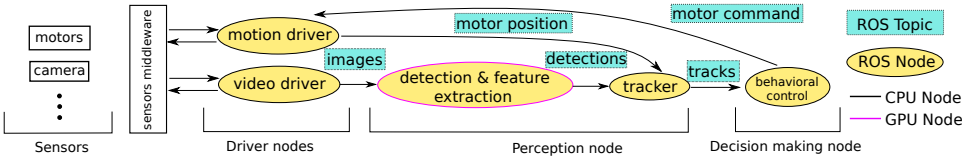


Figure 2.1: The robotic software architecture is composed of several nodes: an image is produced by the video driver, fed to the face detector which produces both face detections and appearance features, then transmitted to the tracker node (alongside motor position information). The tracking results are exploited by the robot control to move the robot's head exploiting the motors drivers.

robot's sensors and produce Topics containing images and motor positions. Camera's topics are processed by the face detection and feature extraction Nodes to produce detections and deep appearance descriptors transmitted to the tracker Node. The tracker Node takes advantage of the motor information and detections to update the appearance model and produce track information following (2.8), transmitted to the behavioral control, which uses it to update the motors' position, depending on the predefined policy.

In the algorithm previously described, the appearance model update is done every few frames. Since this update is time consuming, the tracker cannot wait for the model update to finish before keep on tracking. The appearance model update will run in the background, and the tracker will use the latest appearance model update available.

2.6 EXPERIMENTS

We evaluate our joint tracking and appearance model update robotic architecture in different settings. First, we provide quantitative results following the standard evaluation protocol on the MOT16 dataset. Then, we provide quantitative evaluation under the "active surveillance camera" and the "robot navigation in crowded scenes" application, using MOT17 training set. Finally, we provide qualitative evaluation under the social conversation application.

2.6.1 QUANTITATIVE EVALUATION

Dataset For the sake of reproducibility and in order to be able to compare different tracking systems in the exact same conditions, we use the well-known MOT16 and MOT17 datasets [76] using the public detection setting. The MOT16 dataset is composed of a dozen of videos taken in various conditions, and provides detections obtained with one detector (DPM [30]), while MOT17 use the same video sequences but exploit additional detections obtained with two extra detectors (FR-CNN [87] and SDP [107]). They both provide ground truth on the training set. On MOT17, the standard protocol is to report the results averaged over the three detectors.

Importantly, two kinds of videos are available: recorded with a surveillance camera, and with a camera mounted on an autonomous robot navigating in crowded scenes. Both scenarios are of interest for us. Indeed, the first scenario allows us to simulate the motion of a surveillance camera, and to gauge the robustness of the proposed tracking system against ego-motion noise. The second scenario provides the opposite case in which the ego-motion is completely unknown and must be inferred from visual information.

The first scenario (*moving surveillance camera*) consists on emulating that the surveillance camera only sees half of the image width, and then moving the emulated field of view accordingly to a pre-defined trajectory. The ego-motion vector E_t is then contaminated with Gaussian noise with a standard deviation of η pixels in a uniformly sampled direction. The second scenario (*robot navigating in the crowd*) consists on using the full field of view of the camera, and estimate the ego-motion vector E_t with an optical-flow-based strategy.

Implementation details In all scenarios, the appearance model is updated every 5 frames for 2 epochs. We create the appearance training set by sampling 50 images per identity, and then sample the image pairs as described in Section 2.5.1. Weight decay with a factor $\lambda = 10^{-4}$ is used to regularize the training process. The appearance model CNN is instantiated by a ResNet [43] architecture, where the last layer was replaced by a two layer perceptron with 500 and 100 units activated with ReLu. This CNN is pre-trained for the person re-identification task on the Market-1501 [120] and DukeMTMC [123] datasets, following [126].

Optical flow (OF) is extracted [29] and is used in two different ways. First, in the *robot navigating in the crowd* settings, the OF is used to estimated the ego-motion vector E_t, by

Table 2.1: Results on the MOT16 test set, and standard evaluation setting.

Model	Detection		Tracking		ID	
	Rcll	Prcn	MOTA	MOTP	IDF1	IDsw
CH [7]	**45.4**	87.8	**38.4**	75.4	37.8	1,321
ODA-UP	40.4	**91.6**	36.3	**75.6**	**48.0**	**757**

averaging the OF over the entire image. In both scenarios the average optical flow within one detection bounding box, provides an estimate of the velocity of the track (after the ego-motion vector is subtracted). In order to get stable results we set s_n^2 to a fixed value: once $\phi_{\boldsymbol{w}}$ is trained, we know the that the squared radius of the clusters in that space is $\tau - 1$. We then set the value of s_n^2 accordingly.

Evaluation protocol We compare our *online deep appearance update* (ODA-UP) based method with the state-of-the-art in multi-person tracking for social robotics [7]. While the tracking model is very similar, the appearance model previously used in the literature is based on *color histograms* (CH). We first compare MOT performance on MOT16 test set using the standard evaluation procedure where unmodified sequences are used, as [7] is originally benchmarked on MOT16. We then compare tracking performance on MOT17's training set using the *moving surveillance camera* and *robot navigating in the crowd* scenarios to better assess their performance in a robotic setting. Since the tracking framework on which both trackers are derived from are generative probabilistic models, and are thus not preliminary trained, we can leverage training data and annotations in a fair comparison. Additionally, and in order to provide a full evaluation of the necessity of the on-line appearance model update, we compare the proposed tracker with the exact same architecture without updating the weights of the deep appearance model, and refer to it as ODA-FR, for frozen. In that case, the appearance likelihood is provided by computing the cosine similarity between appearance templates and current detections. For the *moving surveillance camera* scenario, we evaluate under different values of $\eta \in \{0, 0.8, 1.6, 3.2\}$. We report standard multiple object tracking metrics: the recall (Rcll), the precision (Prcn), the number of identity switches (IDsw), the fragmentation (FM), the multiple object tracking accuracy (MOTA) and precision (MOTP). These metrics are well known and were introduced in [11]. The ID consistency of the tracks is further evaluated with identity recall (IDR), precision (IDP) and F1 measure (IDF1), see [89].

Discussion Table 2.1 reports results on MOT16 test set, using the standard evaluation protocol. We note that CH achieves better overall tracking performance (+2.1%), which is mainly explained by a high track recall (+5%). However, our proposed strategy achieves significantly higher performance in terms of tracks identity consistency: +10.2% in IDF1, which translates into nearly half less identity switches (IDsw). ODA-FR is not reported here because MOT16's test server doesn't allow multiple submissions. This higher performance in re-identification is not surprising since our approach uses a more powerful

Table 2.2: Results on MOT17's *moving surveillance camera* setting .

η	Model	Detection		Tracking		Identities		
		Rcll	Prcn	MOTA	MOTP	IDP	IDR	IDF1
0	CH [7]	49.4	88.2	42.5	84.5	70.3	39.4	50.5
	ODA-FR	49.5	**88.7**	43.0	**84.8**	66.7	37.2	47.8
	ODA-UP	**54.7**	86.7	**45.6**	84.0	**75.4**	**45.7**	**56.0**
0.8	CH [7]	49.6	88.0	42.5	84.4	69.9	39.4	50.4
	ODA-FR	49.7	**88.7**	43.1	**84.7**	67.1	37.6	48.2
	ODA-UP	**54.4**	86.3	**45.0**	83.8	**71.2**	**44.9**	**55.1**
1.6	CH [7]	49.1	88.2	42.2	84.2	70.3	39.1	50.2
	ODA-FR	49.5	**88.6**	42.8	**84.5**	66.3	37	47.5
	ODA-UP	**54.5**	86.4	**45.3**	83.7	**73.3**	**46.2**	**56.7**
3.2	CH [7]	49.2	88.2	42.3	83.2	68.1	38.0	48.8
	ODA-FR	49.1	**88.4**	42.4	**83.3**	66.8	37.1	47.7
	ODA-UP	**54.2**	86.1	**44.8**	82.8	**71.5**	**45.0**	**55.2**

Table 2.3: Results on MOT17's *robot navigating in the crowd* settings.

Model	Detection		Tracking		Identities		
	Rcll	Prcn	MOTA	MOTP	IDP	IDR	IDF1
CH [7]	45.8	91.8	41.2	80.7	74.1	37.0	49.3
ODA-FR	45.8	**93.1**	42.0	81.0	73.8	36.3	48.6
ODA-UP	**52.3**	90.5	**46.2**	**81.5**	**79.0**	**45.7**	**57.9**

appearance model, and we argue that in the context of social robotics, re-ID tracking performance is a more relevant measure. We explain our poor performance in MOTA by the relative invariance of this metric to ID switches, and also by the high number of non moving sequences in the dataset, which reduces the risk of track fragmentation and thus interest for a discriminative appearance model, and is therefore not representative of robotic settings. This motivates us to investigate further the impact of a moving robotic head in MOT in the scenarii *moving surveillance camera* and *robot navigating in the crowd*.

Table 2.2 and 2.3 report the results in the two scenarii on MOT17. Regarding the *moving surveillance camera* setting in Table 2.2, we first observe that our approach significantly outperforms both the frozen (FR) and the color histogram (CH) models, by more than +3% and +2% respectively in MOTA. Unsurprisingly, the pretrained appearance model outperforms color histogram based model by roughly +0.5% in MOTA. While different levels of ego-motion noise lead to different scores, the ranking between the methods stays the same. The difference in MOTP is quite small, meaning that the quality of the output bounding boxes (only the tracked ones) is roughly the same. The slight decrease for ODA-UP is due to the fact that ODA-UP is able to track people that are harder to track, and for which estimating good bounding boxes is more challenging. This is sup-

ported by the relative position of the methods in the other metrics. Indeed, the recall and precision metrics are another proof that ODA-UP is able to track significantly more people in the context of a moving camera. Regarding the identity measures, we can see that ODA-UP exhibits by far the highest performance, putting forward the advantage of the adaptive strategy. Indeed, the ODA-UP model outperforms the other two. Interestingly, ODA-FR is outperformed (in identity measures) by the color histograms, demonstrating that complex deep models are useful only if trained in relevant data or, as we propose in this chapter, if they are adapted online.

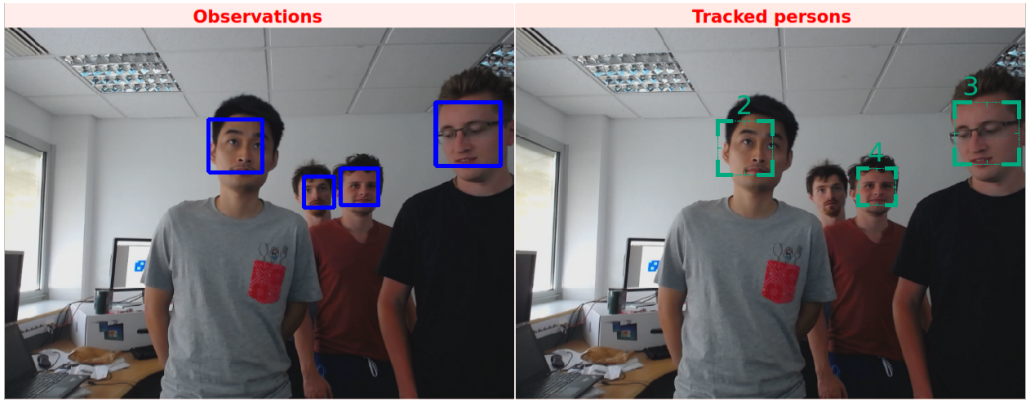
A similar situation is found in Table 2.3 for *robot navigating in the crowd* setting: our approach outperforms the histogram model and the pretrained model by respectively +5.0% and +4.4% in MOTA. The rest of the metrics follow the same ranking as in the previous setting. Very importantly, the findings in Table 2.2 are further confirmed by larger advantage margins in Table 2.3. In both experiments, we observe how that update of the deep appearance model brings two main advantages. First, the tracking recall increases, and thus MOTA does, because leveraging appearance information becomes crucial in the case of a moving camera. Second, and more important, the consistency of the tracks' ID exhibits a significant increase when updating the model online.

2.6.2 QUALITATIVE RESULTS

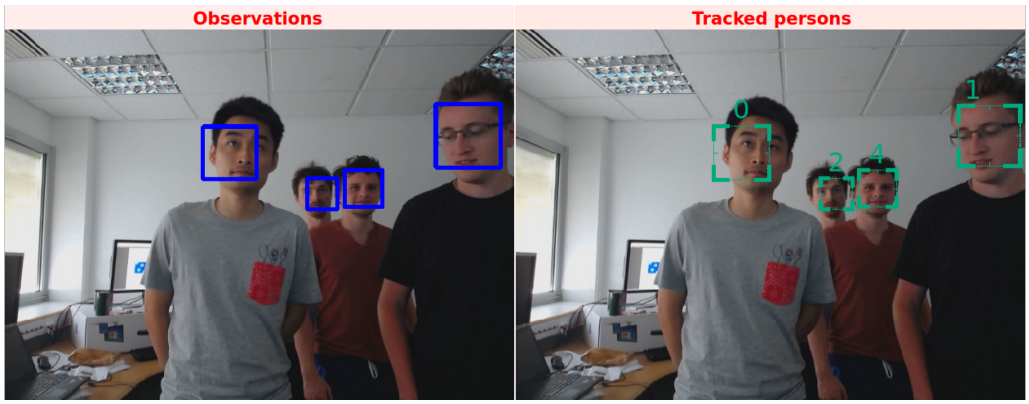
We qualitatively evaluate the performance of the tracker on a real robotic platform, as described in Section 2.5.3, and provide them as videos alongside the chapter: results are available at `https://team.inria.fr/perception/research/oda_track/`. In that case, the procedure in [7] is used to compute the ego-motion vector from the motors velocity. Only faces' bounding boxes are extracted from the images using [1]. Example of the tracking results are displayed in Figure 2.2, using CH and ODA-UP. We note that since our deep metric formulation has a higher discriminative power than color histogram based appearance model, it is able to better distinguish ID 2 and ID 4, even if they are close and that both detections could be generated by a unique ID. Also, we note that the ID labels of the tracks differ significantly when comparing both methods, which is caused by the high number of identity switches in the CH setting.

2.7 CONCLUSION

In this chapter, we address the problem of online multiple object tracking in a domain-agnostic robotic setting, using a joint probabilistic and deep appearance model that allow the update of the appearance embedding simultaneously to tracking multiple people, while accounting for the robot ego-motion. We demonstrate its performance quantitatively using the standard evaluation protocol on MOT16 and in two modified scenarios on MOT17, and qualitatively onboard of a consumer robot.



(a) Tracking result using CH [7]



(b) Tracking result using ODA-UP

Figure 2.2: Tracking qualitative results using CH and ODA-UP. Detections are displayed on the left panel (blue), and tracking results are available on the right panel (green) in 2 settings.

CHAPTER 3

CAMERA ADVERSARIAL UNSUPERVISED RE-ID

3.1 INTRODUCTION

Person re-identification (Re-ID) is a well-studied retrieval task that consists in associating images of the same person across cameras, places and time. Given a query image of a person, we aim to recover his/her identity (ID) from a set of identity-labeled gallery images. The person Re-ID task is particularly challenging for two reasons. First, the query images correspond to IDs never seen before (i.e. during training). Second, the gallery and the query images are captured under a variety of background scenes, illumination conditions, and viewpoints.

Most Re-ID models assume the availability of heavily labeled datasets, and focus on improving their performance on the very same datasets [126, 123, 121, 120, 119, 62]. However, datasets are recorded in specific places and time, and consequently, can be severely biased in terms of background and illumination conditions. These biases partially explain why many Re-ID methods available in the literature generalize poorly to other datasets [28, 25], and as seen in chapter 2. Clearly, this reduces their deployability and usability in real-world scenarios like practical multiple person tracking. In order to overcome these limitations, several methods were recently proposed attempting to address *unsupervised* person Re-ID [99, 112, 28, 25, 83, 57]. These methods assume the availability of a *source* dataset annotated with person IDs and another unlabeled *target* dataset, and aim to optimize the *target* re-ID performance. In that sense, unsupervised Re-ID is formulated as a domain adaptation problem rather than an unsupervised learning problem.

In parallel, and since generative adversarial networks (GANs) [40] were proposed, adversarial learning has gained popularity in the domain adaptation field [97, 34, 13]. The underlying intuition is that learning a feature generator robust to the domain shift

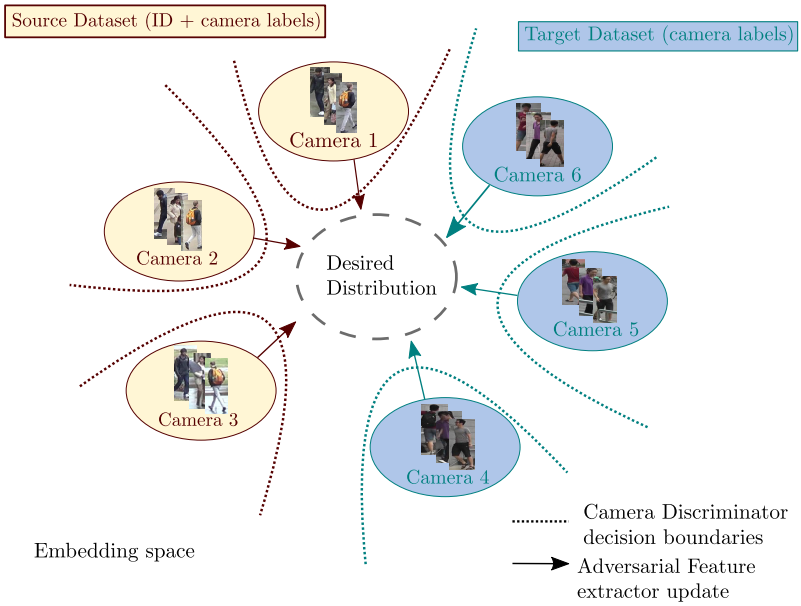


Figure 3.1: We propose an adversarial method to learn a feature extractor defining an embedding space where the different camera distribution match, and are not distinguisible. To achieve that, we extend the classic adversarial domain adaptation method to a multiple domain formulation, by using a camera-classifier as discriminator, instead of the usual binary classifier. At the same time, we ensure that our features are ID-discriminative by using a standard classification loss.

between *source* and *target* would improve the target performance. The adversarial learning paradigm has been successfully used for person Re-ID, mostly for data augmentation, in both the supervised (standard) [126, 123] and the unsupervised (domain adaptation) [25, 70] learning paradigms.

In this chapter, we explore unsupervised re-ID to improve at the pretraining stage the person re-ID model leveraged in chapter 2 as appearance model backbone for MOT, in the case where the target domain is known and unlabeled training data available. We propose a new method exploiting adversarial strategies for person Re-ID. Our approach is summarized in Figure 3.1. The intuition behind the core hypothesis in this chapter is that camera information, i.e. which camera acquired which image, can be used to learn features that are robust to camera changes and, as a consequence, robust to illumination, background and viewpoint changes. This hypothesis can then be used to transfer information from a labeled person Re-ID source dataset to an unsupervised target dataset. Towards this aim, we propose a method referred to as camera adversarial transfer (CAT). We report experiments on the DukeMTMC-ReID [123] and Market-1501 [120] datasets and discuss the impact of different architectural and methodological choices.

3.2 Related Work

3.2.1 Supervised person re-ID

Most Re-ID models are supervised and employ a single dataset for both learning and testing. A plethora of techniques was proposed in order to improve the performance obtained in this standard setting [119, 62, 53, 19]. These methods fall into two categories. First, metric learning based approaches obtain an optimal representation by adequately optimizing the distance between pairs of images [62, 53, 58, 122, 103, 67, 114, 19]. Second, a classifier is trained to return the index of the visible persons [119, 85, 18, 102, 63]. However, independently of the approach employed, these methods perform poorly when the training and the test sets differ significantly in terms of lighting conditions or image quality, as it is the case in cross-dataset experiments [25, 28].

3.2.2 Unsupervised person re-ID

To face the generalization ability issue, unsupervised cross-dataset person Re-ID methods were recently proposed [28, 112, 83, 25, 99, 57]. By jointly learning on a labeled *source* dataset and an unlabeled *target* dataset, these methods exploit the relatively simple process of gathering unlabeled detections from a *target* camera network. For instance, clustering techniques are used in [28, 112, 70] to infer information about the *target* IDs and to incorporate the estimated IDs into the training procedure. In [83, 57], dictionary learning approaches are proposed in order to learn a dataset-shared representation. Interestingly, some other works [99, 94] employ attribute information available with the source dataset, e.g. gender, haircut or clothing style. They show how exploiting these pieces of information improves the performance in the context of unsupervised person Re-ID, at the cost of manually annotating all these attributes. Recently [72, 78] used spatio-temporal constraints on the target camera network to improve the Re-ID performance.

3.2.3 Adversarial for Unsupervised re-ID

Some recent work investigated generative adversarial networks (GANs) for learning person Re-ID as a data augmentation/transformation technique. [123] showed that the use of GAN generated images trained on person Re-ID datasets can improve baseline models. In parallel, an image-to-image translation technique is employed in [126] to augment the training dataset by learning the style of each camera in the supervised Re-ID setting. However, such a strategy suffers from a scalability problem when applied to unsupervised person Re-ID, since the number of learned style translations increases exponentially with the number of target cameras. A similar approach is employed in the framework of unsupervised person Re-ID in [25] in order to transfer a *target* style onto labeled *source* detections: by learning the target's image distribution while preserving the source ID information, they train a generator to match the target data distribution in the image domain and use the transferred images to learn a baseline Re-ID model. Intuitively, our proposed

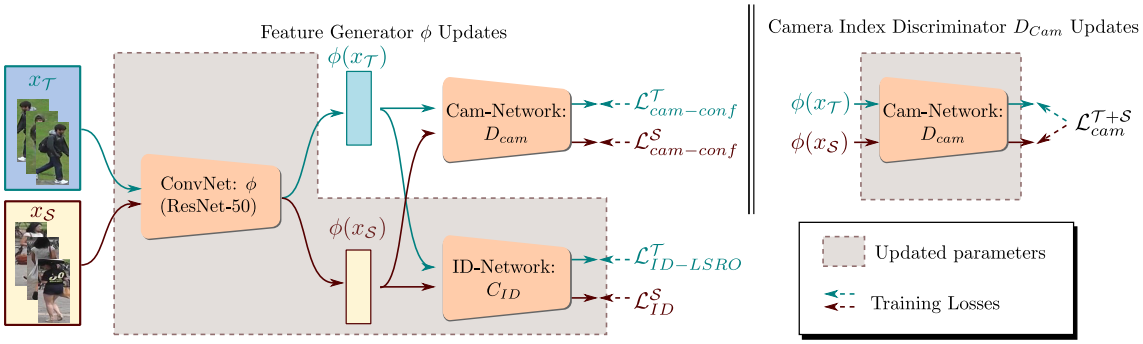


Figure 3.2: Architecture of our camera adversarial transfer strategy. The generator is first updated using source's indentity labels, and minimize the performance of the camera discriminator, using losses (3.8) and (3.9). The discriminator is then updated to recognize from which cameras the feature vectors are extracted with (3.10).

method is similar to this approach: we match domain distributions directly in the feature domain instead of the image domain. In this way, we remove the data generation step, and we unify the learning procedure. More recently, [124] has been proposed to learn each camera style independently and use metric learning approaches to make use of the generated examples and train a Re-ID model.

Adversarial learning for domain adaptation has recently emerged [34, 97, 13]. It aims to train a discriminator to distinguish between *source* and *target* features, such that a feature generator is trained to find a domain-invariant and discriminative representation for the target task. The effectiveness of adversarial domain adaptation for unsupervised person Re-ID has been showed in [34], where a coarse discriminator is trained to distinguish between source and target features, as if they were real and fake samples with a standard GAN.

In order to achieve robustness across camera networks, we propose to cast the problem in the adversarial framework. Considering a standard person Re-ID model, we take advantage of the adversarial loss to learn an embedding space where the different camera feature distributions match, i.e. they are indistinguishable. To do so, we extend the classic adversarial domain adaptation method to a multiple domain formulation, by using a camera-classifier as discriminator, instead of the usual binary classifier. At the same time, we ensure that our learned features are ID-discriminative by using a classification loss. We demonstrate its interest in the context of unsupervised person re-ID with a standard baseline. To the best of our knowledge, ours is the first attempt to exploit adversarial learning to directly tackle the data distributions discrepancies related to camera changes in the context of person Re-ID.

3.3 CAMERA ADVERSARIAL TRANSFER (CAT)

As previously mentioned, the aim of CAT is to understand whether a camera-wise adversarial strategy could be useful in the context of unsupervised person Re-ID. More formally, let $\mathcal{S}$ denote a source ID-annotated person Re-ID dataset, containing $N^{\mathcal{S}}$ images corresponding to $M^{\mathcal{S}}$ different identities captured by $K^{\mathcal{S}}$ cameras. We write $\mathcal{S} = \{(\boldsymbol{x}_n^{\mathcal{S}}, \boldsymbol{p}_n^{\mathcal{S}}, \boldsymbol{c}_n^{\mathcal{S}})\}_{n=1}^{N^{\mathcal{S}}}$, where each three-tuple consists of a detection image, $\boldsymbol{x}_n^{\mathcal{S}}$, a person ID one-hot vector, $\boldsymbol{p}_n^{\mathcal{S}} \in \{0,1\}^{M^{\mathcal{S}}}$ and a camera index one-hot vector, $\boldsymbol{c}_n^{\mathcal{S}} \in \{0,1\}^{K^{\mathcal{S}}}$. Similarly, we define $\mathcal{T} = \{(\boldsymbol{x}_n^{\mathcal{T}}, \boldsymbol{c}_n^{\mathcal{T}})\}_{n=1}^{N^{\mathcal{T}}}$ a target person Re-ID dataset, with $K^{\mathcal{T}}$ cameras and $N^{\mathcal{T}}$ element, without ID labels.

Let ϕ be a convolutional neural network front-end (e.g. ResNet-50 to fix ideas) considered as our *feature extractor*. The goal of person Re-ID is to be able to discriminate between identities, and therefore an identity classifier C_{ID} is required. The cross-entropy loss is usually employed:

$$\mathcal{L}_{\text{ID}}^{\mathcal{S}}(\phi, C_{\text{ID}}) = -\mathbb{E}_{(\boldsymbol{x}^{\mathcal{S}}, \boldsymbol{p}^{\mathcal{S}}) \sim \mathcal{S}} \left\{ \log \left\langle C_{\text{ID}}(\phi(\boldsymbol{x}^{\mathcal{S}})), \boldsymbol{p}^{\mathcal{S}} \right\rangle \right\}, \tag{3.1}$$

where $\mathbb{E}$ denotes the expectation and $< \cdot, \cdot >$ denotes the scalar product, in this case between the output of the ID discriminator $C_{\text{ID}}(\phi(\boldsymbol{x}^{\mathcal{S}}))$ and the ground-truth ID vector $\boldsymbol{p}^{\mathcal{S}}$. Training person Re-ID systems with $\mathcal{L}_{\text{ID}}$ loss alone in a supervised setting has been widely studied and can be considered a well-established methodology [25, 123, 126].

In this context, we investigate how to use camera index information to improve the transferability of the learned visual features to a camera-only supervised dataset. One of the main issues is that the features learned with the classical strategy discussed above describe a combination between the identity of the person and camera specifics (such as background, illumination conditions, viewpoint..). Because in the cross-dataset setting the datasets are recorded with different cameras, it results in a significant discrepancy between the two learned feature distributions, and has a strong negative impact on the performance. Formally, this distribution shift can be measured according to the generalized Jensen-Shannon divergence [68],

$$\underset{1 \leq c \leq K^{\mathcal{S}} + K^{\mathcal{T}}}{JSD} (p(\phi(X)|C = c)) \neq 0 \tag{3.2}$$

This is confirmed empirically: we show that when ϕ is trained using (3.1) only, we can then train a classifier to predict the true camera index from the features of ϕ (see the experimental section).

Recent works in domain adaptation and adversarial learning [40, 34] show that this framework can be used for matching the source's and target's feature distribution. They do so, first by training a dataset classifier on top of the feature extractor, and then adding an adversarial term in the loss. We get inspired from this strategy to propose a multiple domain (camera) discriminator, mixing camera and ID-labels from the source target together with camera-only labels from the target domain, to match camera embedding distribution and minimize the divergeance in Eq. (3.2).

To implement this, we require a camera index discriminator D_{Cam} (see Figure 3.2 for a complete overview of the architecture) and we define a camera index (cross-entropy) loss:

$$\mathcal{L}_{\text{cam}}^{\mathcal{S}+\mathcal{T}}(\phi, D_{\text{cam}}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{c})\sim\mathcal{S}+\mathcal{T}}\left\{\log\left\langle D_{\text{cam}}(\phi(\boldsymbol{x})), \boldsymbol{c}\right\rangle\right\} \tag{3.3}$$

On one side, the feature extractor must minimize the person Re-ID loss $\mathcal{L}_{\text{ID}}$ at the same time as making the problem more challenging for the camera discriminator. On the other side, the camera discriminator tries to learn to identify the camera from which the generated feature $\phi(\boldsymbol{x}^{\mathcal{S}/\mathcal{T}})$ has been extracted :

$$\min_{\phi, C_{\text{ID}}} \max_{D_{\text{cam}}} \mathcal{L}_{\text{ID}}^{\mathcal{S}}(\phi, C_{\text{ID}}) - \mu\mathcal{L}_{\text{cam}}^{\mathcal{S}+\mathcal{T}}(\phi, D_{\text{cam}}), \tag{3.4}$$

where $\mu > 0$ is a parameter with two possible interpretations. From a domain adaptation perspective, it can be seen as a regularization parameter [34]. From an adversarial learning perspective, μ adjusts the impact of the discriminator's gradient when training the generator.

Under mild hypothesis, we can show (Annex A) that minimizing the loss (3.4) is equivalent to the following optimization problem

$$\begin{aligned} &\min_{\phi, C_{\text{ID}}} \mathcal{L}_{\text{ID}}^{\mathcal{S}}(\phi, C_{\text{ID}}) \\ &s.t.\ JSD(p(\phi(x)|c=1), .., p(\phi(x)|c=K^{\mathcal{S}}+K^{\mathcal{T}})) = 0 \end{aligned} \tag{3.5}$$

Although this adversarial loss could directly be applied to our problem, we empirically show its limitations (cf table 3.2 in the experimentation section), and propose methods to alleviate its problems in the following sections.

3.3.1 DOMAIN CONFUSION LOSS

The adversarial loss as formulated in (3.4) suffers from convergence problems [40, 3], since the discriminator converges quickly at early stages of the training leading to vanishing gradient problems. The solution usually adopted [40, 97] is to train the generator with a binary crossentropy loss and inverted labels, considering generated images as if their were sampled from the true distribution.

In addition, in the context of our camera adversarial learning, when updating the generator, the adversarial loss encourages lower probability values for the true camera labels but does not constrain all the other camera label probabilities to simultaneously increase. Consequently, it is likely that the generator favors the visually closest cameras. Since the visually closest camera usually belongs to the same dataset, this standard adversarial formulation would not force feature distribution matching between the source and target datasets.

To alleviate this problem and inspired by [97], we propose a camera confusion loss by imposing that features extracted from ϕ result in a uniform distribution labelisation

according to the camera-discriminator:

$$\mathcal{L}^{\mathcal{R}}_{\text{cam-conf}}(\phi, D_{\text{cam}}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{c})\sim\mathcal{R}}\left\{\log\left\langle D_{\text{cam}}(\phi(\boldsymbol{x})), \boldsymbol{u}\right\rangle\right\}$$
$$\text{where } \mathcal{R} \in \{\mathcal{S}, \mathcal{T}\} \text{ and } \boldsymbol{u} = \frac{1}{K^{\mathcal{S}} + K^{\mathcal{T}}} \tag{3.6}$$

Note that $\mathcal{R}$ is equal to either $\mathcal{S}$ or $\mathcal{T}$ depending on the current training step as detailed in Sec.3.3.3. When updating ϕ according to (3.6), we take into account all the discriminator outputs and therefore encourage the model to match feature distributions across datasets.

It is also close to the alternative adversarial loss described earlier, in the sense that we encourage ϕ to increase all cameras output probabilities instead of only diminishing the true cameras probability output.

3.3.2 Label smoothing regularization for Outliers

The identity distributions across datasets do not have overlapping support (ie the datasets do not share identity labels). Solving the optimization problem (3.5) might end up producing features discriminative in terms of source identities, invariant to camera changes, and to target's identities. This problem is refered to as negative transfer. It would drasticaly hurt target's re-ID performance. To mitigate this problem, we propose to use Label Smoothing Regularization for Outliers (LSRO) [123] for ID supervision of target's images.

$$\mathcal{L}^{\mathcal{T}}_{\text{ID-LSRO}}(\phi, C_{ID}) = -\mathbb{E}_{\boldsymbol{x}^{\mathcal{T}}\sim\mathcal{T}}\left\{\log\left\langle C_{\text{ID}}(\phi(\boldsymbol{x}^{\mathcal{T}})), \boldsymbol{u}_{\boldsymbol{M}^{\mathcal{S}}}\right\rangle\right\}$$
$$\boldsymbol{u}_{\boldsymbol{M}^{\mathcal{S}}} = 1/M^{\mathcal{S}} \tag{3.7}$$

It has two interpretations: similarly to [123], it acts as a regularization term, introducing appearance variations when evaluating the ID loss, preventing the ID classifier to focus on under-represented visual features correlated with identities. Second, it prevents the feature distribution matching to crush target's ID-specific features, by incorporating the information that no target IDs are present in the source dataset, and therefore should lie far for from source's ID decision boundaries, thus mitigating the negative transfer problem.

3.3.3 Overall training

In practice, both the identity classifier and the camera discriminator are implemented as classification layers within the neural architecture. The difference between them lies on how their weights are updated, i.e. with which information are the classifier and the discriminator trained. From the two optimization problems stated above, we can clearly see that, while the feature extractor is updated with both the identity and the camera losses, the ID classifier and the camera discriminator are updated with the person and camera index information respectively. Even if both the classifier and the discriminator are classifying/discriminating, only the second is a discriminator within the adversarial

philosophy. We end up jointly solving the following optimization problems

$$\min_{\phi, C_{\text{ID}}} \mathcal{L}^{\mathcal{S}}(\phi, C_{ID}, D_{cam}) = \mathcal{L}^{\mathcal{S}}_{\text{ID}}(\phi, C_{\text{ID}}) + \mu \mathcal{L}^{\mathcal{S}}_{\text{cam-conf}}(\phi, D_{\text{cam}}), \tag{3.8}$$

$$\min_{\phi, C_{\text{ID}}} \mathcal{L}^{\mathcal{T}}(\phi, C_{ID}, D_{cam}) = \lambda \mathcal{L}^{\mathcal{T}}_{\text{ID-LSRO}}(\phi, C_{\text{ID}}) + \mu \mathcal{L}^{\mathcal{T}}_{\text{cam-conf}}(\phi, D_{\text{cam}}), \tag{3.9}$$

$$\min_{D_{\text{cam}}} \mathcal{L}^{\mathcal{S}+\mathcal{T}}(\phi, C_{ID}, D_{cam}) = \mathcal{L}^{\mathcal{T}+\mathcal{S}}_{\text{cam}}(\phi, D_{\text{cam}}), \tag{3.10}$$

where μ and λ are relevant only when training ϕ.

We optimize those losses iteratively, alternating between the three. We call our method Camera Adversarial Transfer (CAT).

3.4 EXPERIMENTS

3.4.1 PROTOCOL

Datasets. The proposed adversarial transfer strategies are evaluated mainly on two datasets: Market-1501 (Market) [120] and DukeMTMC-reID (Duke) [89, 123]. In both cases, the dataset consists on three parts: a training set, a gallery set and a query set. The query and gallery subparts are never available during training, and only used for testing purposes. The Market dataset is composed of $M = 1,501$ (half for training and half for test) different identities, observed through $K = 6$ different cameras (viewpoints). The deformable parts model [30] is used to detect people in the images. As a consequence we obtain $N = 12,936$ images for training and $19,732$ for the gallery images. The query subparts consists of $3,368$ hand-drawn bounding boxes. The Duke dataset is composed of $M = 1,404$ (half for training and half for test) identities captured from $K = 8$ different cameras. In addition, 408 other ID, called "distractors", are added to the gallery. Detections are manually selected, leading to $N = 16,522$ images for train, $17,661$ for the gallery and $2,228$ queries. Transfer experiments trained on Market and evaluated on Duke are noted Market $\rightarrow$ Duke, and reversely Duke $\rightarrow$Market.

Evaluation metrics. In order to provide an objective evaluation of the performance of the adversarial strategies, we employ two classical metrics in person Re-ID [120]: Rank-1 (R1) and mean average-precision (mAP). In more details, for each query image, we extract the visual features employing the adversarially learned generator ϕ, and we match them to the features extracted from the images of the gallery using cosine distance. Importantly, the features corresponding to the gallery images captured with the same camera as the query image are not considered. For R1, a query is well identified if the closest gallery feature vector corresponds to the same identity. In the case of mAP, the whole list of gallery images (except those taken with the same camera) is considered, and the precision of those rank positions in which a positive match is found is averaged. See [120] for details. For both metrics, the mean over the query set is reported.

Implementation and training details. The feature extractor ϕ used in our study corresponds to a recent variant [126] of the well know IDE [121] and denoted by IDE+. IDE is often used as baseline to conduct experimental evaluations for the task of person Re-ID, see [121, 126, 123, 25, 28, 110]. Both architectures are based on ResNet-50, which we implicitly assume pre-trained on ImageNet unless otherwise specified.

In detail, The IDE+ baseline consists in adding the ID classifier after the last pooling layer of ResNet-50 (usually denoted as *pool5*). The ID classifier consists of a hidden fully connected layer of 1,024 units, a batch normalization layer activated with ReLU, a dropout regularization layer (with rate at 0.5), and a fully connected classification layer activated by a softmax function. The number of units of the classification layer depends on the number of training person identities. The overall architecture is trained end-to-end using Stochastic Gradient Descent (SGD).

The camera discriminator consists of one hidden layer of 256 units each, with ReLU activations regularized with weight decay (with rate at 0.01). The camera classification layer is activated with softmax and the number of units depends on the number of cameras of the dataset. An ablation study on the number of units per layer is presented later on in this experimental section. Besides, we faced a pratical issue when training the discriminator: it is not a balanced classification problem. In other words, the number of detections per camera strongly depends on the camera, and therefore if we do not carefully address this issue, we will be biasing the discriminator towards those cameras in which more people are detected. In order to tackle this problem, we sample evenly across cameras when gathering a batch.

We resize the input images to 256×128 pixels to keep the aspect ratio. We use mirroring and a batch size of 64 (for the sake of the stability of the results). The pre-training is inspired from [126]. The model is initialized on ImageNet and pre-trained for 20 epochs in the supervised setting (using only source information). The camera matching training is then conducted for 10 epochs, using the same optimization settings. We use a fixed learning rate for all the experiments ($\eta = 0.01$), and use a learning rate multiplier when training the ID branch, which multiplies by 10 the ID classifier's gradient.

For inference, we use local max pooling (LMP), introduced in [25], which consists in removing the final global average layer of ResNet, and replace it with 2 local max pooling, one pooling from the half-top of the picture, the other pooling the other part. We therefore obtain a feature vector of dimension 2×2048.

3.4.2 Comparison to State of the Art

We compare our results with two hand-crafted feature-based classical strategies: Bag of Words (BoW) [120] and local maximal occurence (LOMO) [67]. We also compared against clustering and finetuning (PUL) [28] as well as transferable joint attribute-identity deep learning (TJ-AIDL) [99]. Importantly, this method exploits extra labeled data (i.e. attributes) which must be manually annotated and that the rest of the methods do not employ. Results of Bottom-Up Clustering Approach [70] (BUC) are also reported. We compare to the closest work in the literature: SPGAN [25] and HHL [124].

Table 3.1: R1 and mAP measures on both datasets. The first part of the table is extracted from the literature. All the results are obtained under the same experimental protocol. † refers to unsupervised settings exploiting extra annotations (attributes). Overall best results are shown in **bold**, second best results are in ***italic***.

Method	Duke → Market		Market→ Duke	
	R1	mAP	R1	mAP
LOMO [67]	27.2	8.0	12.3	4.8
BoW [120]	35.8	14.8	17.1	8.3
PUL [28]	45.5	20.5	30.0	16.4
TJ-AIDL [99] †	57.1	26.2	39.6	22.0
TJ-AIDL [99] †	58.2	26.5	44.3	23.0
SPGAN [25]	51.5	22.8	41.1	22.3
SPGAN + LMP [25]	58.1	26.9	46.9	26.4
HHL [124]	***62.2***	***31.4***	46.9	27.2
BUC [70]	**66.2**	**38.3**	***47.4***	***27.5***
IDE+	45.6	19.8	32.4	16.8
IDE+ + LMP	53.5	25.4	41.2	23.2
CAT + LSRO	54.9	26.2	45.9	24.6
CAT + LSRO + LMP	57.8	27.8	**50.9**	**28.7**

All these results are copied directly from the literature. Table 3.1 reports the results when using Duke and Market as source and target (left) and vice versa (right).

From Table 3.1 we can see that deep learning based methods (including ours) outperform hand-crafted features, confirming once more the interest of deep learning for unsupervised person Re-ID. We see that our method yields competitive results compared to the state of the art, getting the best results for the Market → Duke (+3.5% in R1, and + 1.2% in mAP), and third best for Duke →Market. We also note that our approach is significantly simpler than SPGAN [25] and HHL [124] (our closest competitive methods) to train, since those methods require first to train a GAN inspired generator to fit target's image distribution, generate a transfered dataset and then train the baseline on top of it. In a practical point of view, our method is therefore more practical and scalable, since it only requires a few epochs of training from a pretrained network to get competitive results on a new unlabeled dataset.

3.4.3 Ablation study

Adversarial learning variants We use the adversarial framework to train the model with different training settings. We use the adversarial loss (3.4) that we denote by *adv*, to compare with our domain confusion objective (see section 3.3.1). It best performs for $\mu = 0.01$ and 5 epochs.

We also experiment the classic adversarial domain adaptation setting, when the dis-

criminator is only trained to recognize source's and target's features. We call this variant Domain Adversarial Transfer (DAT). Since the task of the discriminator is easier than in the CAT setting, we use a simpler architecture (1 layer, 64 neurons and weight decay at 0.1) to prevent the model to overfit on identities. The best global performances are reported, for $\mu = 0.01$ and 5 epochs.

Finally, we experiment the adversarial setting exploiting only source's information: the discriminator is only fed source's detections, and ϕ is updated using only source's image, ID and camera labels. We refer to this method as Fully Adversarial Training (FAT). When training this method, we use a 2 layer discriminator of 512 neurons and train it from scratch (pretrained imagenet weights), and train it for 20 epochs. The results of those different variants are reported in table 3.2. Results with LMP are included for the sake of completeness, but not considered in the following analysis.

All experiments demonstrate asymetric performances when switching dataset's role. It generally performs better when considering Duke as target (+13.5% vs +10.3% in best scenarios). An explanation might come from the fact that ID information is closely linked to the camera index information in the Duke dataset, in comparison to the Market. This is backed by the mutual information measure between identities and camera index in each domain: 0.84 for Duke, and 0.14 for Market. It means that providing camera index information will give the model much more identity information in the Duke than in the Market. Therefore, it is more difficult for the model to learn features not dependent on cameras specifics (background/illumination) while being ID-discriminative, thus making the features less transferable. It is also confirmed by robustness experiments, developed in the next section.

Second, FAT slightly improves the Re-ID performance compared to our baseline in both cases (+2.7% and 3.8% respectively), without exploiting target information. It confirms the interest for adversarial learning in the context of domain generalization [64].

Interestingly, CAT experiments yield better results than DAT (+7% and +5%). Although DAT experiments use a simpler discriminator, and explicitly enforce a strong regularization, we show that using a labelization less dependent on identities for the discriminator, like camera index, encourages the discriminator to exploit features less ID dependent, and thus yield better re-ID performance.

Finally, we demonstrate the interest of the camera confusion loss in the context of person Re-ID (+3.1%,+0.9%) and of LSRO (+2.8%,+2.7%).

Robustness assessment The robustness of the features learned can be easily evaluated, to have a sense of how well camera's feature distributions match. We split the Gallery (test set) of each dataset into a gallery-training set and a gallery-test set (2/3-1/3). We then train a camera-discriminator (same implementation as in 3.4.1). We evaluate its camera accuracy performance (the lower the better). Note that the dataset split is done making sure that there is no ID overlap between the 2 gallery sub-datasets, to ensure that we do not use ID-dependent information during the camera classification task.

Table 3.2: Adversarial strategies variants. FAT uses only source data at training with $\mathcal{L}_{\text{cam-conf}}$, DAT's discriminator is supervised only with dataset's labels, and *adv* refers to a standard adversarial loss.

# exp. setting	Duke→Market		Market→Duke	
	R1	mAP	R1	mAP
IDE+	45.6	19.8	32.4	16.8
FAT	48.3	20.9	36.2	19.3
DAT	45.1	19.6	37.3	19.8
CAT + *adv*	49.0	21.6	42.3	22.2
CAT	52.1	24.3	43.2	23.6
CAT + LSRO	***54.9***	***26.2***	***45.9***	***24.6***
CAT + LSRO + LMP	**57.8**	**27.8**	**50.9**	**28.7**

Table 3.3: Feature's robustness evaluation: we extract features from the gallery images of a given dataset, split them into a testing and training set, and train a camera classifier with the latter on top of the frozen evaluated model. We report the camera accuracy performance on the testing set of each re-ID dataset. Better camera robustness results are in **bold**. Refer to the text for detail.

# exp. setting	Camera classifier's accuracy			
	Duke→Market		Market→Duke	
IDE+	85.5	67.6	68.2	91.0
FAT	**45.5**	62.0	**42.7**	88.0
DAT	77.6	64.4	62.8	84.4
CAT + *adv*	56.4	49.1	48.1	**68.1**
CAT	61.0	**27.9**	45.4	68.6
CAT + LSRO	71.0	33.2	50.1	68.2

First of all, we notice that baseline (IDE+) robustness performance is limited, since we manage to train a camera-classifier with good accuracy. It has been the preliminary experiment motivating our adversarial approach. The ID-supervised datasets have better robustness measures, meaning that a classic supervised technique already enforce some kind of robustness, which explain their good performance in a supervised setting.

Second, the Duke dataset consistently has higher accuracy scores across experiments. The camera robustness is tougher to achieve for this dataset, and it comforts our strong ID-camera relationship hypothesis explaining the poorer performance in Duke→Market.

All adversarial experiments encourage camera robustness for both source and target datasets. It is noticeable that FAT achieves it without having access to target information. CAT does a better job than CAT+*adv* in Duke → Market, which might explain why it outperforms it significantly in Re-ID performance (see table 3.2). We also note that using LSRO increases the accuracy performance of the camera classifier, indicating that its use mitigates the adversarial strategy impact in order to preserve ID-related information.

Table 3.4: CAT + LSRO R1 and mAP performance for different number of neurons for the discriminator and μ. We note acc the accuracy performance of the camera discriminator at the end of the training.

# units	Duke→Market			Market→Duke		
	R1	mAP	acc	R1	mAP	acc
128	52.9	24.1	35.9	44.7	23.9	**30.1**
256	**54.9**	**26.2**	**31.4**	**45.9**	**24.6**	33.2
512	50.0	22.7	43.9	43.4	23.0	32.6
1,024	51.7	23.7	33.3	42.8	23.0	34.6
μ	Duke→Market			Market→Duke		
	R1	mAP	acc	R1	mAP	acc
0.01	45.0	19.4	64.3	41.4	22.2	62.3
0.05	48.5	21.7	44.1	45.5	24.4	40.3
0.1	**54.9**	**26.2**	31.4	**45.9**	24.6	33.2
0.2	52.9	24.8	**26.5**	45.3	**24.7**	**25.6**

Impact of camera variability. We train CAT with different target dataset sizes to understand the impact of the variability of the target dataset on the Re-ID performance. The only modification on the experimental protocol is that the target dataset consists on what is captured from cameras 1 to $\tilde{K}$, making $\tilde{K}$ vary from 1 (only one camera) to $K^{\mathcal{T}}$ (all cameras). In these experiments we use Market as source dataset, and Duke as target. The gain in performance over the IDE+ baseline is shown in Figure 3.3.

We first notice that the IDE+ baseline is respectively better, equivalent and worse when using one, two or more than two cameras. Generally, we observe that the Re-ID performance regularly increases with the number of cameras. The regular increase in performance when adding more and more cameras is a clear trend in Figure 3.3. Our understanding is that the adversarial strategy is good at capturing and exploiting the intra-dataset variability, and satisfactorily exploits the different viewpoints to learn more robust and discriminative person Re-ID features. CAT do not only learn to match feature distributions across datasets, but also takes full advantage of the target camera network by matching the different camera distributions.

Impact of the hyper-parameter μ and discriminator architecture. We evaluate the impact of the value of μ on the overall performance of the system in table 3.4. First of all, we observe that the Re-ID performance is relatively stable when changing the training parameters, although the Duke $\rightarrow$ Market is less stable, more specifically when μ is too low. The Re-ID performance first increases with the adversarial loss weight, and then decrease when it is too strong. Considering the final accuracy of the discriminator, it is conform to intuition. The stronger the adversarial weight gets, the less accurate our discriminator becomes. When comparing the different architectures, we see that an optimal discriminator is found for 258 neurons, and that the Re-ID performance is quite robust to a change of architecture.

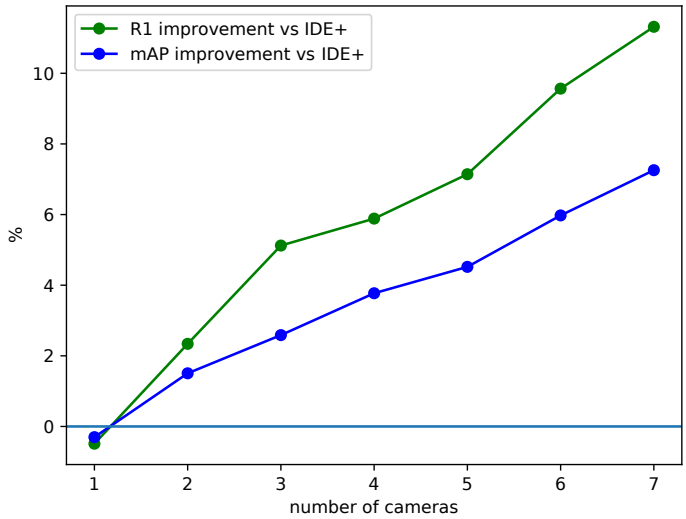
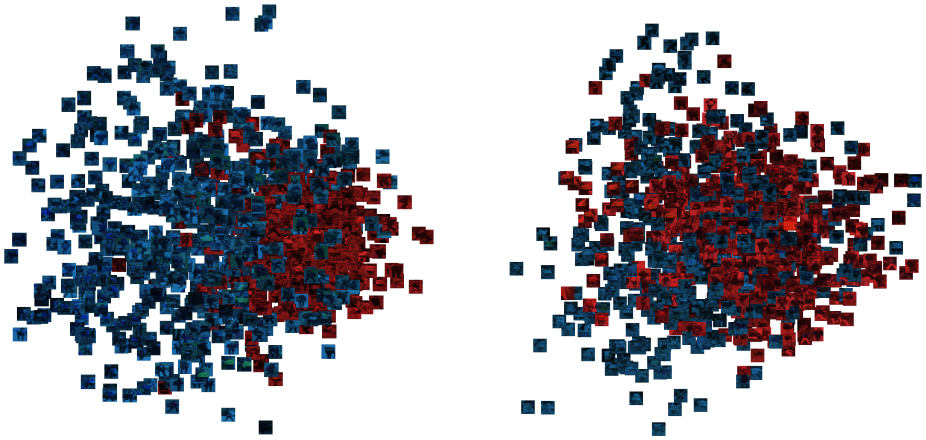


Figure 3.3: Performance variation with the number of cameras in the setting Market →Duke

Embedding visualization We use the feature embedding to compute a PCA projection of the learned space, and use the first 2 dimensions for visualization. In Figure 3.4 both source and target datasets are used to compute the projection, while in Figure 3.5, 3.6, 3.7, 3.8, and 3.9, only the target dataset is used. We use PCA projection to preserve global structure, which is not guaranteed in other dimensionality reduction methods like t-SNE. The Market → Duke training setting is used for both, and we use the training sets of each respective datasets. In dataset-wise visualization, we observe that in the IDE+ setting, using only source information, both distributions visually differs. It fits our strong dataset-shift hypothesis, and motivates our approach. The features learned with CAT have a distribution discrepancy significantly lower, which indicates that our approach actively helps to match the 2 domain distributions. Similarly, in the camera-wise visualization, we observe a stronger separation between cameras when trained with IDE+ than when trained with our adversarial approach. This analysis is backed by the accuracy measures reported in table 3.3, where we show that we can't train a generalizable camera-classifier, comforting the fact that feature distributions are less separated when learned with CAT. It visually confirms that a camera-based approach improves the results not only because it pushes source and target distribution together, but also matches each camera distribution, especially target's cameras.

3.5 Conclusion

In this chapter we address the problem of unsupervised person Re-ID in the setting of unsupervised domain adaptation. More precisely, we propose to use camera index information within an adversarial paradigm that we name *camera adversarial transfer*. The

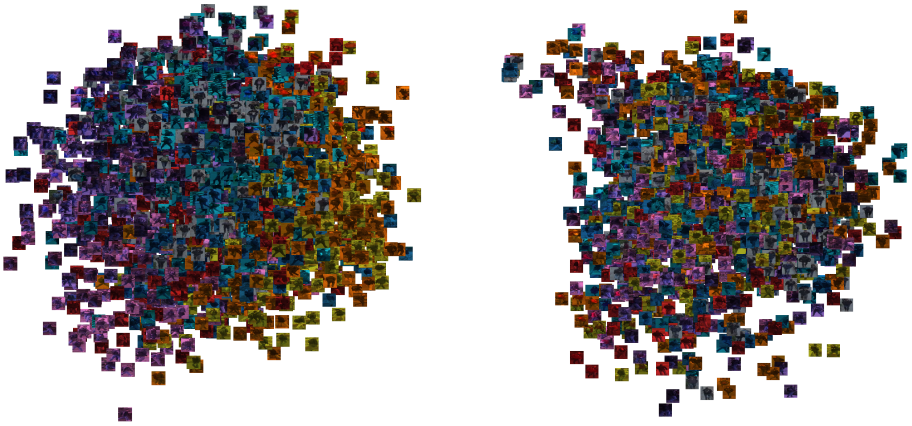


(a) PCA projection of $\mathcal{S} + \mathcal{T}$ learned with IDE+.

(b) PCA projection of $\mathcal{S} + \mathcal{T}$ learned with CAT.

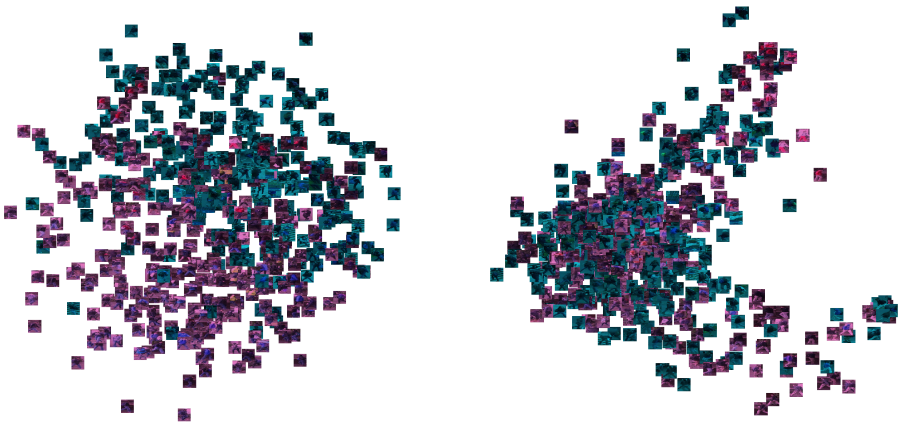
Figure 3.4: Source dataset is in red and target is in blue. Best viewed in color.

proposed strategy are compared to several baselines and adversarial variants inspired from the domain adaptation literature, on the Duke→Market and Market→Duke experimental settings, thus demonstrating the interest and benefits of exploiting camera index information within an adversarial framework for person Re-ID. We also evaluate these strategies in terms of camera robustness, discuss the influence of the structure of the camera discriminator as well as the number of cameras and the hyper-parameter μ.



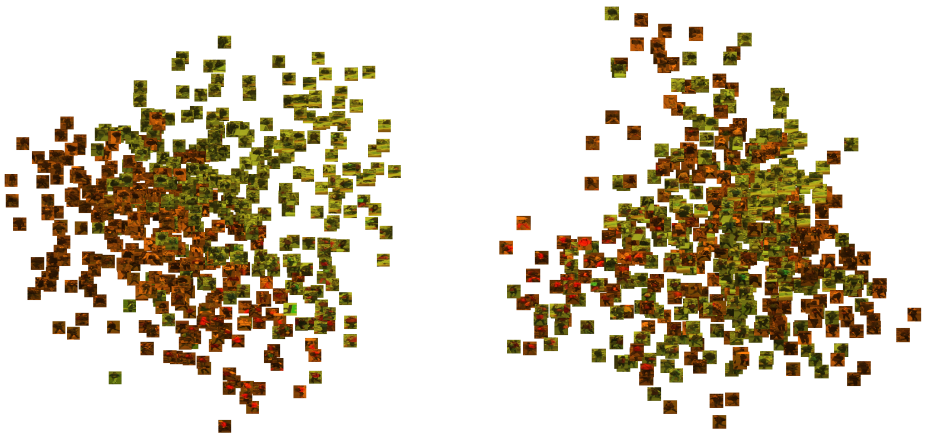
(a) PCA of $\mathcal{T}$ with IDE+ baseline. (b) PCA of $\mathcal{T}$ with CAT (Ours).

Figure 3.5: The PCA is only done with all target's cameras.



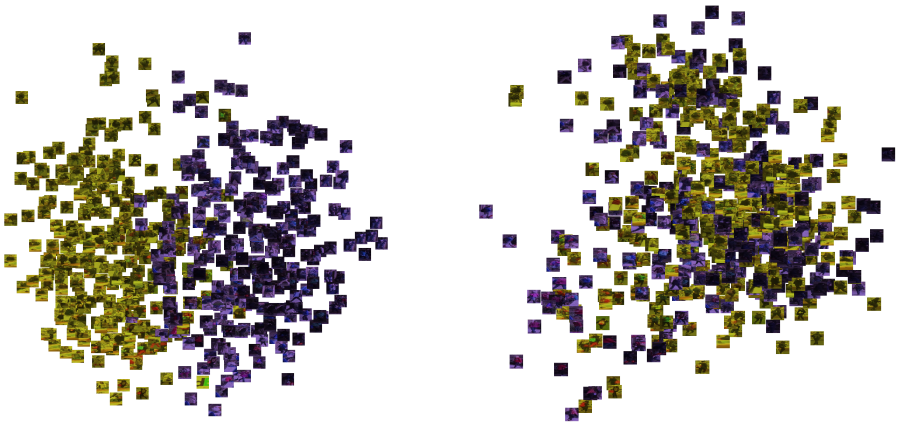
(a) PCA projection of $\mathcal{T}$ with IDE+. (b) PCA projection of $\mathcal{T}$ with CAT.

Figure 3.6: The PCA is only done with target's 7th (purple) and 8th (blue) camera. Best viewed in color.



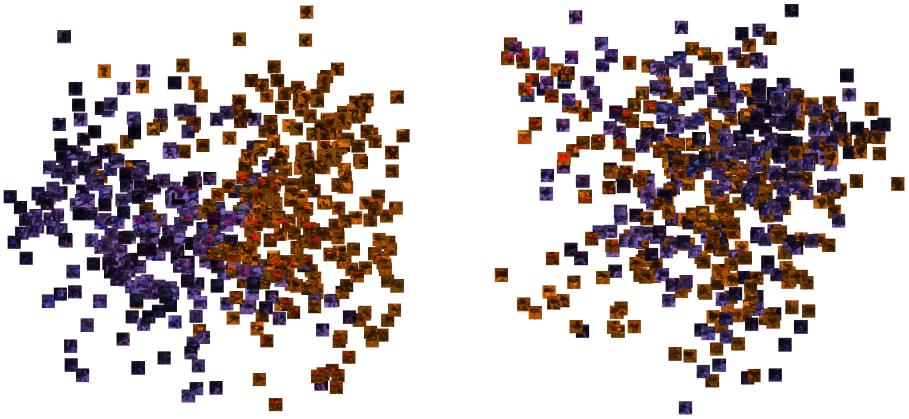
(a) PCA of $\mathcal{T}$ with IDE+ baseline. (b) PCA of $\mathcal{T}$ with CAT (Ours).

Figure 3.7: The PCA is only done with 1st and 2nd target's cameras.



(a) PCA of $\mathcal{T}$ with IDE+ baseline. (b) PCA of $\mathcal{T}$ with CAT (Ours).

Figure 3.8: The PCA is only done with 1st and 3rd target's cameras.



(a) PCA of $\mathcal{T}$ with IDE+ baseline. (b) PCA of $\mathcal{T}$ with CAT (Ours).

Figure 3.9: The PCA is only done with 2nd and 3rd target's cameras.

CHAPTER 4

CONDITIONAL ADVERSARIAL NETWORK FOR UNSUPERVISED RE-ID

4.1 INTRODUCTION

As seen in the previous chapter, person re-identification (re-ID) is a well-studied retrieval task [111, 74, 52] that consists in associating images of the same person across cameras, places and time. Most re-ID models assume the availability of labeled datasets and focus on improving their performance on the very same dataset, see for instance [89, 120]. The limited generalization capabilities of such methods [25, 28] lead researchers to overcome this limitation by investigating a new person re-ID task, where there is a *source* dataset annotated with person IDs and another unlabeled *target* dataset. This is called *unsupervised* person re-ID. Contrary to the framework investigated in chapter 3, we investigate the *clustering* and *finetuning* strategy: this recent trend uses a pretrained architecture to extract visual features, cluster them, and use the cluster assignments as *pseudo-labels* to re-train the base architecture using standard supervised re-ID loss functions [32, 35].

In parallel, and as seen in chapter 3, adversarial learning has gained popularity in the domain adaptation field [97, 34, 13]. The underlying intuition is that learning a feature generator robust to the domain shift between *source* and *target* would improve the target performance. The adversarial learning paradigm has been successfully used for person re-ID in both the supervised [126, 123], and, as investigated previously, the unsupervised [28, 70] learning paradigms.

In this chapter, we propose to unify these two trends in unsupervised person re-ID. Naturally, one would expect that an adversarial game between a generator (feature extractor) and a discriminator (camera classifier) should suffice, as seen previously. However, because the ID presence is not uniform in all cameras, such simple strategy implies some negative transfer and limits – often decreases – the representational power of the visual feature extractor. In this chapter, we aim to investigate the impact of Negative Transfer in Unsupervised Person re-ID, and propose to use conditional adversarial networks to address it, by providing an additional identity representation to the camera discriminator.

Since in the target dataset, the ID labels are unavailable, we exploit the pseudo-labels. More precisely, we provide, as conditioning vector, the centroid of the cluster to which the image belongs. The contributions of this chapter are the following:

- We investigate the impact of a camera-adversarial strategy in the unsupervised person re-ID task within the *clustering* and *finetuning* framework.
- We realize the negative transfer effect, and propose to use conditional adversarial networks.
- The proposed method can be easily plugged into any unsupervised clustering-based person re-ID methods. We experimentally combine **CANU** with two clustering-based unsupervised person re-ID methods, and propose to use their cluster centroids as conditioning labels.
- Finally, we perform an extensive experimental validation on four different unsupervised re-ID experimental settings and outperform current state-of-the-art methods by a large margin on all settings.

The rest of the chapter is organized as follows. Section 4.2 describes the state-of-the-art. Section 4.3 discusses the basics of clustering-based unsupervised person re-ID and sets the notations. The proposed conditional adversarial strategy is presented in Section 4.4. The extensive experimental validation is discussed in Section 4.5 before drawing the conclusions in Section 4.6.

4.2 RELATED WORK

4.2.1 UNSUPERVISED PERSON RE-IDENTIFICATION

Unsupervised person re-identification (re-ID) has drawn growing attention in the last few years, taking advantage of the recent achievements of supervised person re-ID models, without requiring an expansive and tedious labeling process of the target data set. A very important line of research starts from a pre-trained model on the source data set and is based on *clustering* and *fine-tuning* [28, 70, 32, 35, 115]. It alternates between a clustering step generating noisy pseudo-labels, and a fine-tuning step adapting the network to the target data set distribution, leading to a progressive label refinement. Thus, these methods do not use the source data set during adaptation. A lot of effort has been invested in improving the quality of the pseudo-labels. Sampling from reliable clusters during adaptation [28], gradually reducing the number of clusters and merging by exploiting intrinsic inter-ID diversity and intra-ID similarity [70], or performing multiple clustering on visual sub-domains and enforcing consistency [32] have been investigated. More recently, [35] investigated the interaction of two different models to assess and incorporate pseudo-label reliability within a teacher-student framework.

A different approach is directly inspired by Unsupervised Domain Adaptation (UDA) [25, 124, 16, 84, 93, 125]: using both the source and target data sets during adaptation. These

methods aim to match the distributions on the two data sets while keeping its discriminative ability leveraging source ground truth ID labels. A first strategy learns to map source's detections to target's style detections, and train a re-ID model in a supervised setting using those only those transferred detections [25], or in combination with the original target detections [124]. More standard UDA strategies use adversarial learning to match the source and target distributions [34, 84].

4.2.2 NEGATIVE TRANSFER

Negative Transfer has been investigated in unsupervised domain adaptation [96], especially for Partial Domain Adaptation (PDA) [118, 15, 113], where target labels are only a subset of the source's. Negative transfer is defined as the inability of an adaptation method to find underlying common representation between data sets and is generally caused by the gap between the distributions of the two data sets being too wide [100] for the algorithm to transfer knowledge. Weighting mechanisms are generally employed to remove the impact of source's outliers class on the adaptation process, either for the matching part [113, 15, 108], the classification part [100], or both [118]. Interestingly, [100] uses a domain discriminator conditioned by source label to perform conditional distribution matching. Investigating negative transfer is not limited to UDA settings. For example, a similar method has been proposed for domain generalization [64], implementing a conditional discriminator to match conditioned domain distributions. By doing so, the impact of the difference between prior label distributions on the discriminative ability of the model is alleviated.

Within the task of unsupervised person re-ID, different cameras could be considered as different domains, and standard matching strategies could be used. However, they would inevitably induce negative transfer as described before for generic domain adaptation. Direct application of PDA methods into the person re-ID tasks is neither simple nor expected to be successful. The main reason is that, while PDA methods handle a few dozens of classes, standard re-ID data sets contain a few thousands of IDs. This change of scale requires a different strategy, and we propose to use conditional adversarial networks, with a conditioning label that describes the average sample in the cluster, rather than representing the cluster index. In conclusion, different from clustering and fine-tuning unsupervised person re-ID methods, we propose to exploit (conditional) adversarial networks to learn visual features that are camera independent and thus more robust to appear changes. Different from previous domain adaptation methods, we propose to match domains (cameras) with a conditioning label that evolves during training, since it is the centroid of the cluster to which the visual sample is assigned, allowing us having a representation that is independent of the number of clusters and the cluster index.

4.3 CLUSTERING BASED UNSUPERVISED PERSON RE-ID

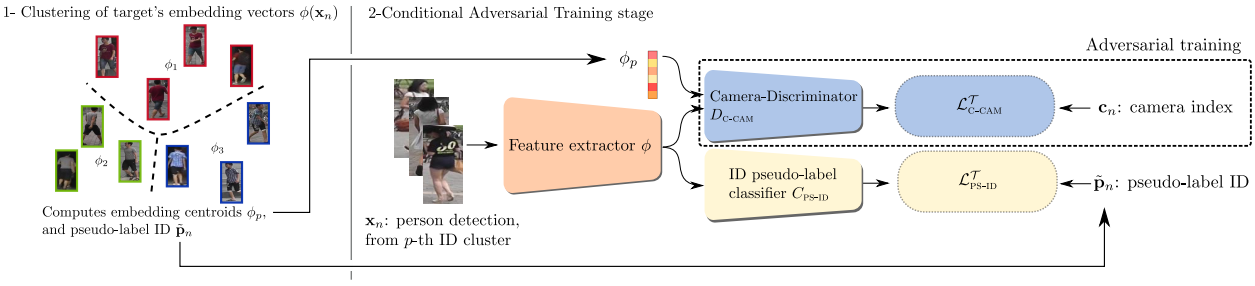


Figure 4.1: Pipeline of our method: alternatively (1) clustering target's training data set using ϕ representation, producing noisy pseudo-label ID $\tilde{p}_n$ alongside centroids ϕ_p, and (2) conditional adversarial training, using a Camera-Discriminator D_{CAM} conditioned by ϕ_p to enforce camera invariance on a per identity basis to avoid negative transfer. Pseudo-label ID are used to train an ID classifier C_{PS-ID} alongside the discriminator.

We propose to combine conditional adversarial networks with the *clustering* and *fine-tuning* framework for unsupervised person Re-ID. To detail our contributions, we first set up the basics and notations of existing methods for unsupervised person re-ID.

Let $\mathcal{S}$ denote a source ID-annotated person re-ID dataset, containing $N^{\mathcal{S}}$ images corresponding to $M^{\mathcal{S}}$ different person identities captured by $K^{\mathcal{S}}$ cameras. We write $\mathcal{S} = \{(\boldsymbol{x}_n^{\mathcal{S}}, \boldsymbol{p}_n^{\mathcal{S}}, \boldsymbol{c}_n^{\mathcal{S}})\}_{n=1}^{N^{\mathcal{S}}}$, where each three-tuple consists of a detection image, $\boldsymbol{x}_n^{\mathcal{S}}$, a person ID one-hot vector, $\boldsymbol{p}_n^{\mathcal{S}} \in \{0,1\}^{M^{\mathcal{S}}}$ and a camera index one-hot vector, $\boldsymbol{c}_n^{\mathcal{S}} \in \{0,1\}^{K^{\mathcal{S}}}$. Similarly, we define $\mathcal{T} = \{(\boldsymbol{x}_n^{\mathcal{T}}, \boldsymbol{c}_n^{\mathcal{T}})\}_{n=1}^{N^{\mathcal{T}}}$ a target person re-ID dataset, with $K^{\mathcal{T}}$ cameras and $N^{\mathcal{T}}$ element, without ID labels.

Source pre-training Let ϕ be a convolutional neural network backbone (*e.g.* ResNet-50 [43]) served as a trainable *feature extractor*. The goal of person re-ID is to be able to discriminate person identities, and therefore an identity classifier C_{ID} is required. The output of C_{ID} is a $M^{\mathcal{S}}$-dimensional stochastic vector, encoding the probability of the input to belong to each of the identities. The cross-entropy and triplet losses are usually employed:

$$\mathcal{L}_{\text{CE}}^{\mathcal{S}}(\phi, C_{\text{ID}}) = -\mathbb{E}_{(\boldsymbol{x}^{\mathcal{S}}, \boldsymbol{p}^{\mathcal{S}})\sim\mathcal{S}} \left\{ \log \left\langle C_{\text{ID}}(\phi(\boldsymbol{x}^{\mathcal{S}})), \boldsymbol{p}^{\mathcal{S}} \right\rangle \right\}, \tag{4.1}$$

$$\begin{aligned} \mathcal{L}_{\text{TRI}}^{\mathcal{S}}(\phi) =& \mathbb{E}_{(\boldsymbol{x}^{\mathcal{S}}, \boldsymbol{x}_p^{\mathcal{S}}, \boldsymbol{x}_n^{\mathcal{S}})\sim\mathcal{P}_{\mathcal{S}}} \{\max(0, \|\phi(\boldsymbol{x}^{\mathcal{S}}) - \phi(\boldsymbol{x}_p^{\mathcal{S}})\| \\ &+ m - \|\phi(\boldsymbol{x}^{\mathcal{S}}) - \phi(\boldsymbol{x}_n^{\mathcal{S}})\|)\}, \end{aligned} \tag{4.2}$$

where $\mathbb{E}$ denotes the expectation, $\langle \cdot, \cdot \rangle$ the scalar product, $\|.\|$ the L^2-norm distance, $\boldsymbol{x}_p^{\mathcal{S}}$ and $\boldsymbol{x}_n^{\mathcal{S}}$ are the hardest positive and negative example for $\boldsymbol{x}^{\mathcal{S}}$ in $\mathcal{P}_{\mathcal{S}}$ the set of all triplets in $\mathcal{S}$, and $m = 0.5$. We similarly denote $\mathcal{L}_{\text{CE}}^{\mathcal{T}}$ and $\mathcal{L}_{\text{TRI}}^{\mathcal{T}}$ the cross-entropy and triplet losses evaluated on the target dataset. However, in unsupervised reID settings, target ID labels are unavailable, and therefore we will need to use alternative *pseudo-ID labels*. The re-ID

feature extractor ϕ is typically trained using:

$$\mathcal{L}_{\text{ID}}^{\mathcal{S}}(\phi, C_{\text{ID}}) = \mathcal{L}_{\text{CE}}^{\mathcal{S}}(\phi, C_{\text{ID}}) + \lambda \mathcal{L}_{\text{TRI}}^{\mathcal{S}}(\phi), \tag{4.3}$$

for a fixed balancing value λ, achieving competitive performance on the source test set [44]. However, they notoriously lack generalization power and perform badly on datasets unseen during training [25], thus requiring adaptation.

Target fine-tuning As discussed above, target ID labels are unavailable. To overcome this while leveraging the discriminative power of widely-used losses described in Eq. 4.3, methods like [32, 35] use pseudo-labels. The hypothesis of these methods is that the features learned during the pre-training stage are exploitable for the inference of target's ID labels to a certain extent. Starting from the pre-trained model, these methods alternate between (i) pseudo ID label generation $\{\tilde{\boldsymbol{p}}_n^{\mathcal{T}}\}_{n=1}^{N^{\mathcal{T}}}$ using a standard clustering algorithm (k-means or DBSCAN [27]) on the target training set $\{\phi(\boldsymbol{x}_n^{\mathcal{T}})\}_{n=1}^{N^{\mathcal{T}}}$ and (ii) the update of ϕ using losses similar to Eq. 4.3 supervised by $\{\tilde{\boldsymbol{p}}_n^{\mathcal{T}}\}_{n=1}^{N^{\mathcal{T}}}$. Since our approach is agnostic to the ID loss used at this step, we choose to denote it by $\mathcal{L}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}})$, $C_{\text{PS-ID}}$ being an optional classifier layer for the pseudo-labels, and develop it further in the experimental section.

4.4 CANU-ReID: A Conditional Adversarial Network for Unsupervised Person re-ID

In this section we discuss the main limitation of clustering-based unsupervised re-ID methods: we hypothesize that viewpoint variability can make things difficult for clustering methods and propose two alternatives. First, an adversarial network architecture targeting re-ID features that are camera-independent. This strategy could, however, induce some negative transfer when the correlation between cameras and IDs is strong. Second, a conditional adversarial network architecture specifically designed to overcome this negative transfer.

Camera adversarial-guided clustering We hypothesize that camera (viewpoint) variability is one of the major limiting factors for clustering-based unsupervised re-ID methods. In plain, if the embedding space variance explained by camera changes is high, the clustering method could be clustering images from the same camera, rather than images from the same ID. Therefore, ϕ will produce features that can very well discriminate the camera at the expense of the ID. To alleviate this problem, we propose to directly enforce camera invariance in ϕ's representation by using an adversarial strategy, where the discriminator is trained to recognize the camera used to capture the image. Consequently, the generator, in our case ϕ, is trained to remove any trace from the camera index (denoted by $\mathbf{c}$). Intuitively, this should reduce the viewpoint variance in the embedding space, improve pseudo-labels quality and increase the generalization ability of ϕ to unseen IDs.

To do so, we require a camera discriminator D_{CAM} (see Fig. 4.1 for a complete overview of the architecture). The generator ϕ and the discriminator D_{CAM} will be trained through a min-max formulation:

$$\min_{\phi, C_{\text{PS-ID}}} \max_{D_{\text{CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) - \mu \mathcal{L}^{\mathcal{T}}_{\text{CAM}}(\phi, D_{\text{CAM}}), \tag{4.4}$$

where $\mu > 0$ is a balance hyper-parameter that can be interpreted as a regularization parameter [34], and $\mathcal{L}^{\mathcal{T}}_{\text{CAM}}$ is defined via the cross-entropy loss:

$$\mathcal{L}^{\mathcal{T}}_{\text{CAM}}(\phi, D_{\text{CAM}}) = -\mathbb{E}_{(\boldsymbol{x}^{\mathcal{T}}, \boldsymbol{c}^{\mathcal{T}}) \sim \mathcal{T}} \left\{ \log \left\langle D_{\text{CAM}}(\phi(\boldsymbol{x}^{\mathcal{T}})), \boldsymbol{c}^{\mathcal{T}} \right\rangle \right\} \tag{4.5}$$

On one side, the feature extractor ϕ must minimize the person re-ID loss $\mathcal{L}_{\text{PS-ID}}$ at the same time as making the problem more challenging for the camera discriminator. On the other side, the camera discriminator tries to learn to recognize the camera corresponding to the input image.

Adversarial negative transfer It has been shown [64] that minimizing (4.4) is equivalent to the following problem:

$$\begin{aligned} &\min_{\phi, C_{\text{PS-ID}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) \\ &\quad \text{s.t. } \text{JSD}_{\mathcal{T}}(p(\phi(\boldsymbol{x})|\boldsymbol{c}=1), \dots, p(\phi(\boldsymbol{x})|\boldsymbol{c}=K)) = 0, \end{aligned} \tag{4.6}$$

where $\text{JSD}_{\mathcal{T}}$ stands for the multi-distribution Jensen-Shanon divergence [68] on the target set $\mathcal{T}$, and we drop the superscript $\mathcal{T}$ in the variables to ease the reading.

Since the distribution of ID labels may strongly depend on the camera, the plain adversarial strategy in (4.6) can introduce negative transfer [100]. Formally, since we have:

$$p(\boldsymbol{p}|\boldsymbol{c}=i) \neq p(\boldsymbol{p}|\boldsymbol{c}=j), i \neq j$$

then solving (4.6) is not equivalent (see [64]) to:

$$\begin{aligned} &\min_{\phi, C_{\text{PS-ID}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) \\ &\quad \text{s.t. } \text{JSD}_{\mathcal{T}}(p(\phi(\boldsymbol{x})|\boldsymbol{p}, \boldsymbol{c}=1), \dots, p(\phi(\boldsymbol{x})|\boldsymbol{p}, \boldsymbol{c}=K)) = 0, \end{aligned} \tag{4.7}$$

which is the problem we would implicitly want to solve. Intuitively, *negative transfer* means that the camera discriminator learns $p(\boldsymbol{c}|\boldsymbol{p})$ instead of $p(\boldsymbol{c}|\boldsymbol{x}, \boldsymbol{p})$, exploiting ID to infer camera information and decreasing the representation power of ϕ due to the adversarial loss.

Conditional adversarial networks We propose to directly solve the optimization problem in Eq. 4.7 to alleviate the negative transfer. Similar to the original conditional GAN formulation [79], we condition the adversarial discriminator with the input ID $\boldsymbol{p}$. Given that ID labels are unavailable on the target set, we replace them by the pseudo-labels obtained during the clustering phase.

However, since we are handling a large number of IDs (700 to 1500 in standard re-ID datasets), using a one-hot representation turned out to be very ineffective. Indeed, such representation is not permutation-invariant, meaning that if the clusters are re-ordered, the associated conditional vector changes, which does not make sense. We, therefore, need a permutation-invariant conditioning label.

To do so, we propose to use the cluster centroids $\boldsymbol{\phi_p}$ which are provided by the clustering algorithms at no extra cost. This conditioning vectors are permutation invariant. Importantly, we do not back-propagate the adversarial loss through the ID-branch, to avoid using an ID-dependant gradient from the adversarial loss. This boils down to defining $\mathcal{L}_{\text{C-CAM}}$ as:

$$\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi, D_{\text{C-CAM}}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{p},\boldsymbol{c})\sim\mathcal{T}}\left\{\log\left\langle D_{\text{C-CAM}}(\phi(\boldsymbol{x}), \boldsymbol{\phi_p}), \boldsymbol{c}\right\rangle\right\} \tag{4.8}$$

and then solving:

$$\min_{\phi, C_{\text{PS-ID}}} \max_{D_{\text{C-CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{PS-ID}}(\phi, C_{\text{PS-ID}}) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi, D_{\text{C-CAM}}). \tag{4.9}$$

4.5 Experimental Validation

In this section, we provide implementation details and an in-depth evaluation of the proposed methodology, setting the new state-of-the-art in four different unsupervised person re-ID experimental settings. We also provide an ablation study and insights on why conditional adversarial networks outperform existing approaches.

4.5.1 Evaluation Protocol

We first describe here the baselines, on which our proposed **CANU** is built and tested. The used datasets and the evaluation metrics are then introduced.

Baselines The proposed **CANU** can be easily plugged into any clustering-based unsupervised person re-ID methods. Here, we experimentally test it on two state-of-the-art clustering-based unsupervised person re-ID methods, as baselines.

First, self-similarity grouping [32] (**SSG**) performs independent clustering on the upper-, lower- and full-body features, denoted as ϕ^{U}, ϕ^{L} and ϕ^{F}. They are extracted from three global average pooling layers of the convolutional feature map of ResNet-50 [43]. The underlying hypothesis is that noisy global pseudo-label generation can be improved by using multiple, but related clustering results, and enforcing consistency between them. The triplet loss is used to train the overall architecture.

To implement **CANU**-SSG, we define three different camera discriminators, one for each embedding, $D^{\text{U}}_{\text{C-CAM}}$, $D^{\text{L}}_{\text{C-CAM}}$ and $D^{\text{F}}_{\text{C-CAM}}$ respectively, each fed with samples from the related representation and conditioned by the global embedding ϕ^{F}. In the particular case of **CANU**-SSG, the generic optimisation problem in Eq. 4.9 instantiates as:

$$\min_{\phi} \max_{D^{\text{U,L,F}}_{\text{C-CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{SSG}}(\phi) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^{\text{U}}, D^{\text{U}}_{\text{C-CAM}}) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^{\text{L}}, D^{\text{L}}_{\text{C-CAM}}) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^{\text{F}}, D^{\text{F}}_{\text{C-CAM}}). \tag{4.10}$$

Second, Mutual Mean-Teaching [35] (**MMT**) reduces pseudo-label noise by using a combination of hard and soft assignment: using hard labeling reduces the amount of information given to the model, and using soft labeling allows the cluster's confidence to be taken into account. MMT defines two different models $(\phi^1, C^1_{\text{PS-ID}})$ and $(\phi^2, C^2_{\text{PS-ID}})$, both implemented with a IBN-ResNet-50 [82] backbone, initialized with two different pre-trainings on the source dataset. They are then jointly trained using pseudo labels as hard assignments, and inspired by teacher-student methods, using their own pseudo ID predictions as soft pseudo-labels to supervise each other. Soft versions of cross-entropy and triplet loss are used.

To implement **CANU**-MMT, similar to **CANU**-SSG, we define two camera discriminators $D^1_{\text{C-CAM}}$ and $D^2_{\text{C-CAM}}$, each dedicated to one embedding, and train it using the following instantiation of the generic optimisation problem in Eq. 4.9:

$$\min_{\phi^{1,2}, C^{1,2}_{\text{PS-ID}}} \max_{D^{1,2}_{\text{C-CAM}}} \mathcal{L}^{\mathcal{T}}_{\text{MMT}}(\phi^1, C^1_{\text{PS-ID}}) + \mathcal{L}^{\mathcal{T}}_{\text{MMT}}(\phi^2, C^2_{\text{PS-ID}}) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^1, D^1_{\text{C-CAM}}) - \mu\mathcal{L}^{\mathcal{T}}_{\text{C-CAM}}(\phi^2, D^2_{\text{C-CAM}}). \tag{4.11}$$

While the clustering strategy used in SSG is DBSCAN [27], the one used in MMT is standard k-means. For a fair comparison, we implemented **CANU** with DBSCAN, which has the advantage of automatically selecting the number of clusters. We also evaluate the performance of MMT using the DBSCAN clustering strategy without **CANU**, to evaluate the impact of our method on a fair basis.

Datasets The proposed adversarial strategies are evaluated using three datasets: Market-1501 (Mkt) [120], DukeMTMC-reID (Duke) [89] and MSMT17 (MSMT) [101]. In all three cases, the dataset is divided into three parts: training, gallery, and query. The query and the gallery are never available during training and only used for testing.

Mkt is composed of $M = 1,501$ (half for training and half for testing) different identities, observed through $K = 6$ different cameras (viewpoints). The deformable parts model [30] is used for person detection. As a consequence, there are $N = 12,936$ training images and $19,732$ gallery images. In addition, there are $3,368$ hand-drawn bounding box queries.

Duke is composed of $M = 1,404$ (half for training and half for testing) identities captured from $K = 8$ cameras. In addition, 408 other ID, called "distractors", are added to the gallery. Detections are manually selected, leading to $N = 16,522$ images for train, $17,661$ for the gallery and $2,228$ queries.

MSMT is the largest and most competitive dataset available, with $M = 4,101$ identities ($1,041$ for training, and $3,060$ for test), $K = 15$ cameras, with $N = 32,621$ images for training, $82,161$ for the Gallery and $11,659$ queries.

The unsupervised person re-ID experimental setting using dataset A as source and dataset B as the target is denoted by A ▶ B. We compare the proposed methodology in four different settings: Mkt ▶ Duke, Duke ▶ Mkt, Mkt ▶ MSMT and Duke ▶ MSMT.

Evaluation metrics In order to provide an objective evaluation of the performance, we employ two standard metrics in person re-ID [120]: Rank-1 (R1) and mean average-precision (mAP). Precisely, for each query image, we extract visual features employing ϕ, and we compare them to the features extracted from the gallery using the cosine distance. Importantly, the gallery images captured with the same camera as the query image are not considered. For R1, a query is well identified if the closest gallery feature vector corresponds to the same identity. In the case of mAP, the whole list of gallery images is considered, and precision at different ranking positions is averaged. See [120] for details. For both metrics, the mean over the query set is reported.

Implementation details For both MMT and SSG, we use the models pre-trained on the source datasets (e.g. For Mkt ▶Duke, we use the model pre-trained on the Market dataset and provided by [32] and [35]). DBSCAN is used at the beginning of each training epoch, the parameters for DBSCAN are the same described as in [32]. The weight for (conditional) adversarial losses μ is set to 0.1 for MMT and to 0.05 for SSG, chosen according to a grid search with values between $[0.01, 1.8]$ (see below). The used conditional discriminator has two input branches, one as the (conditional) ID branch and the other is the camera branch, both consist of four fully-connected layers, of size $[2048, 1024]$, $[2048, 1024]$, $[1024, 1024]$, $[1024,$ number of cameras$]$, respectively. Batch normalization [51] and ReLU activation are used. For MMT, during the unsupervised learning, we train the IBN-ResNet-50 [82] feature extractor with Adam [56] optimizer using a learning rate of 0.00035. As default in [35], the network is trained for 40 epochs but with fewer iterations per epoch (400 v.s. 800 iterations) while keeping a similar or better performance. For SSG, we train the ResNet-50 [43] with SGD optimizer using a learning rate of $6e$-5. At each epoch, unlike MMT, we iterate through the whole training set instead of training with a fix number of iterations.

After training, the discriminator is discarded and only the feature extractor is kept for evaluations. For SSG, first, it combines the features extracted from the original image and the horizontally flipped image with a simple sum. Second, the summed features are normalized by their L_2 norm. Finally, The full-, upper- and, lower-body normalized features are concatenated to form the final features. For MMT, the features extracted from the feature extractor are directly used for evaluations.

In the following, we first compare the proposed methodology with the state-of-the-art (see Sec. 4.5.2). Secondly, we discuss the benefit of using conditional camera-adversarial training in the ablation study (see Sec. 4.5.3), and include several insights on the performance of **CANU**.

Table 4.1: Comparison of the proposed **CANU** methodology on the Mkt ► Duke and Duke ► Mkt unsupervised person re-ID settings. **CANU-MMT** establishes a new state-of-the-art in both settings, and **CANU-SGG** outperforms **SSG**.

Method	Mkt ► Duke		Duke ► Mkt	
	R1	mAP	R1	mAP
PUL [28]	30.0	16.4	45.5	20.5
TJ-AIDL [99]	44.3	23.0	58.2	26.5
SPGAN [25]	41.1	22.3	51.5	22.8
HHL[124]	46.9	27.2	62.2	31.4
CFSM [16]	49.8	27.3	61.2	28.3
BUC [70]	47.4	27.5	66.2	38.3
ARN [66]	60.2	33.4	70.3	39.4
UDAP [93]	68.4	49.0	75.8	53.7
ENC [125]	63.3	40.4	75.1	43.0
UCDA-CCE [84]	47.7	31.0	60.4	30.9
PDA-Net [65]	63.2	45.1	75.2	47.6
PCB-PAST [115]	72.4	54.3	78.4	54.6
Co-teaching [42]	77.6	61.7	87.8	71.7
SSG [32]	73.0	53.4	80.0	58.3
CANU-SSG (ours)	76.1	57.0	83.3	61.9
MMT [35]	81.8	68.7	91.1	74.5
MMT (DBSCAN)	80.2	67.2	91.7	79.3
CANU-MMT (ours)	**83.3**	**70.3**	**94.2**	**83.0**

4.5.2 COMPARISON WITH THE STATE-OF-THE-ART

We compare **CANU**-SSG and **CANU**-MMT to the state-of-the-art methods and we demonstrate in Tables 4.1 and 4.2 that **CANU**-MMT sets a new state-of-the-art result compared to the existing unsupervised person re-ID methods by a large margin. In addition, **CANU**-SSG outperforms SSG in all settings. Since the MSMT dataset is more recent, fewer comparisons are available in the experiments involving this dataset, hence the two different tables.

More precisely, the proposed **CANU** significantly improves the performance of the baselines, SSG [32] and MMT [35]. In Mkt ►Duke and Duke ►Mkt (Table 4.1), **CANU**-SSG improves SSG by ↑3.1%/↑3.6% (R1/mAP, same in the following.) and ↑3.3%/↑3.6% respectively, and **CANU**-MMT significantly outperforms MMT by ↑1.5%/↑1.6% and ↑3.1%/↑8.5% respectively. Moreover, for the more challenging setting (Table 4.2), the improvement brought by **CANU** is even more evident. For SSG, for example, we increase the R1/mAP by ↑13.9%/↑5.9% in Mkt ►MSMT, and by ↑11.1%/↑4.6% in Duke ► MSMT. For MMT, **CANU**-MMT outperforms MMT by ↑7.3%/↑8.0% in Mkt ►MSMT, and by ↑8.7%/↑9.0% in Duke ►MSMT. Finally, the consistent improvement in the four

Table 4.2: Comparison of the proposed **CANU** methodology on the Mkt ▶ MSMT and Duke ▶ MSMT unsupervised person re-ID settings. **CANU-MMT** establishes a new state-of-the-art in both settings, and **CANU-SGG** outperforms **SSG**.

Method	Mkt ▶ MSMT		Duke ▶ MSMT	
	R1	mAP	R1	mAP
PTGAN [101]	10.2	2.9	11.8	3.3
ENC [125]	25.3	8.5	30.2	10.2
SSG [32]	31.6	13.2	32.2	13.3
CANU-SSG (ours)	45.5	19.1	43.3	17.9
MMT [35]	54.4	26.6	58.2	29.3
MMT (DBSCAN)	51.6	26.6	59.0	32.0
CANU-MMT (ours)	**61.7**	**34.6**	**66.9**	**38.3**

settings of **CANU**-MMT over MMT (DBSCAN) and the inconsistent improvement of MMT (DBSCAN) over standard MMT proves that the increase of the performance is due to the proposed methodology. To summarize, we greatly improve the baselines using the proposed **CANU**. More importantly, to our best knowledge, we outperform the existing methods by a large margin and establish a new state-of-the-art result.

4.5.3 Ablation Study

In this section, we first perform a study to evaluate the impact of the value of μ. Secondly, we demonstrate the interest of the conditional strategy, versus its non-conditional counterpart. Thirdly, we study the evolution of the mutual information between ground-truth camera indexes and pseudo-labels using MTT (DBSCAN), thus providing some insights on the quality of the pseudo-labels and the impact of the conditional strategy on it. Finally, we visualize the evolution of the number of lost person identities at each training epoch, to assess the impact of the variability of the training set.

Selection of μ We ablate the value μ by comparing the performance (R1 and mAP) of models trained within the range $[0.01, 1.8]$. From Tab. 4.3, $\mu = 0.1$ (**CANU**-MMT) and $\mu = 0.05$ (**CANU**-SSG) yield the best person re-ID performance.

Is conditional necessary? From Table 4.4, we show that the camera adversarial network can help the person re-ID networks trained with clustering-based unsupervised methods better capture the person identity features: **CANU** and adding a simple adversarial discriminator (+Adv.) significantly outperform the baseline methods in all settings. This is due to the combination of the camera adversarial network with unsupervised clustering-based methods. By doing so, the camera dependency is removed from the features of each person thus increasing the quality of the overall clustering. However, because of the negative transfer effect, the camera adversarial network cannot fully exploit the camera information while discarding the person ID information. For this reason, the proposed

Table 4.3: Impact of μ in the performance of **CANU**. When the mAP values are equal, we highlight the one corresponding to higher R1.

Method	μ	Mkt ▶ Duke		Duke ▶ Mkt	
		R1	mAP	R1	mAP
CANU-SSG	0.01	72.8	53.3	79.7	57.2
	0.05	**76.1**	**57.0**	**83.3**	**61.9**
	0.1	74.7	56.2	82.7	61.1
	0.2	75.3	56.5	81.8	60.3
	0.4	73.3	53.5	80.4	59.2
	1.8	7.1	2.9	39.1	17.1
CANU-MMT	0.01	81.3	68.9	92.6	79.2
	0.05	82.4	70.3	93.0	81.3
	0.1	**83.3**	**70.3**	**94.2**	**83.0**
	0.2	82.7	70.3	93.4	82.5
	0.4	82.5	70.3	93.8	82.0
	1.8	82.8	69.9	93.1	81.3

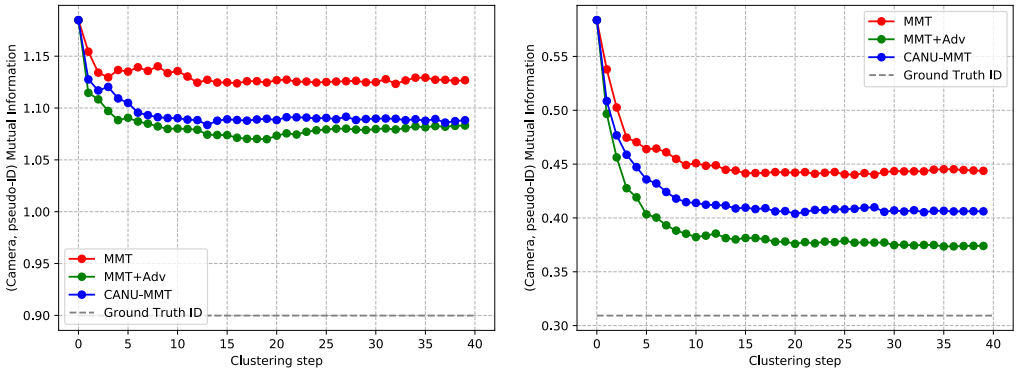
method **CANU** improves the capacity of the camera adversarial network over the simple adversarial strategy. In summary, we demonstrate that the camera adversarial network can help improve the results of unsupervised clustering-based person re-ID. Moreover, the proposed **CANU** further improves the results by removing the link between camera and IDs.

4.5.4 Impact of CANU on Camera Information

Camera information on CANU-MMT Table 4.4 demonstrates that removing camera information is globally positive, but that can also be harmful if it is not done with care. In this section, we further demonstrate that the proposed adversarial strategies actually reduce the camera dependency in clustering results and present some insights on why the conditional strategy is better than the plain adversarial network. To do so, we plot the mutual information between the pseudo-labels provided by DBSCAN, and the fixed camera index information, at each clustering stage (i.e. training epoch) in Fig. 4.2. Intuitively, the mutual information between two variables is a measure of mutual dependence between them: the higher it is, the more predictable one is from knowing the other. We report the results for MMT on Duke ▶ Mkt and Mkt ▶ Duke, **CANU**-MMT and the simple adversarial strategy. We observe that the mutual information is systematically decreasing with the training, even for plain MMT. Both adversarial strategies significantly outperform plain MMT at reducing the camera-pseudo-ID dependency, **CANU**-MMT being slightly less effective than MMT+Adv. This is consistent with our theoretical framework, since matching ID-conditioned camera distribution in ϕ does not account for the ID-Camera dependency, and thus is less effective in terms of camera dependency, but preserves iden-

Table 4.4: Evaluation of the impact of the conditional strategy on SGG [32] and MMT [35] (using DSCAN). When the mAP values are equal, we highlight the one corresponding to higher R1.

Method	Mkt ▶ Duke		Duke ▶ Mkt	
	R1	mAP	R1	mAP
SSG [32]	73.0	53.4	80.0	58.3
SSG+Adv.	75.4	56.4	**83.8**	**62.7**
CANU-SSG	**76.1**	**57.0**	83.3	61.9
MMT (DBSCAN)	80.2	67.2	91.7	79.3
MMT+Adv.	82.6	70.3	93.6	82.2
CANU-MMT	**83.3**	**70.3**	**94.2**	**83.0**

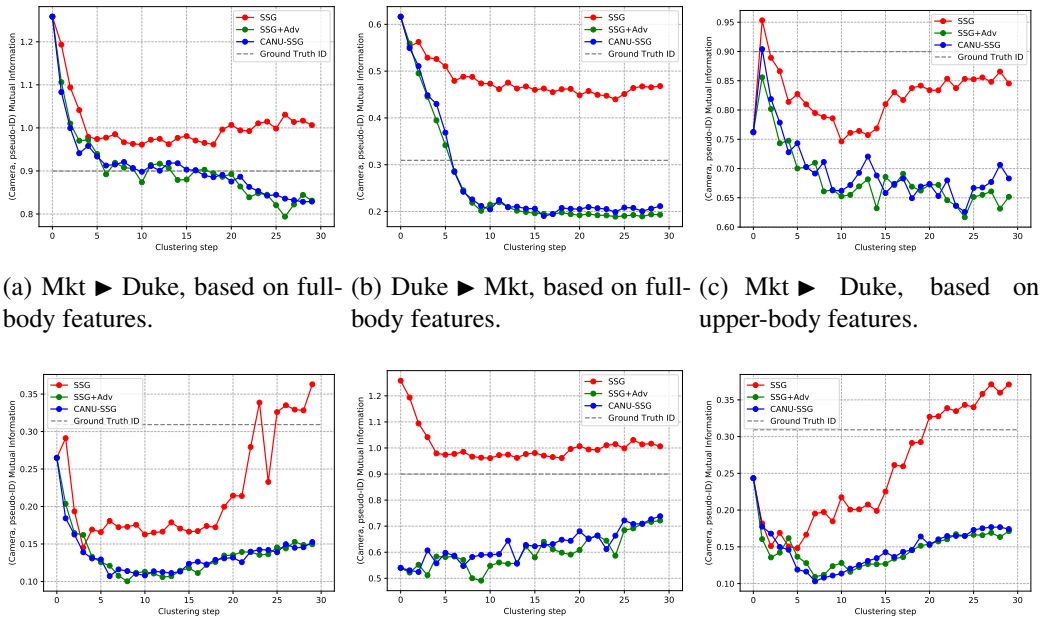


(a) Mkt ▶ Duke (b) Duke ▶ Mkt

Figure 4.2: Mutual information between pseudo labels and camera index evolution for the MMT setting. Ground-truth ID comparison is displayed in dashed lines for both datasets.

tity information, see Table 4.4. We also observe that there is a significant gap between the target mutual information (i.e. measured between ground truth ID and camera index) for all methods, which exhibits the performance gap between supervised and unsupervised person re-ID methods.

Camera information on CANU-SSG From Fig. 4.3, we report the evolution of mutual information between the pseudo-labels provided by DBSCAN [27], and the fixed camera index information with SSG [32] over the clustering stages (i.e. training epochs). Similar to MMT, based on Duke ▶ Mkt and Mkt ▶ Duke settings, we compare SSG, **CANU**-SSG and, the simple adversarial strategy (+Adv). However, since SSG exploits different embedding spaces (i.e. features) and generates a clustering result for each one of them independently, we report the results for full-, upper-, and lower-body features. Moreover, we report target mutual information between ground-truth IDs and camera index. Similar to MMT [35], we observe that plain SSG handles pseudo-ID labels significantly more dependant on camera labels than both adversarial methods, and for all representations. We



(a) Mkt ▶ Duke, based on full-body features.

(b) Duke ▶ Mkt, based on full-body features.

(c) Mkt ▶ Duke, based on upper-body features.

(d) Duke ▶ Mkt, based on upper-body features.

(e) Mkt ▶ Duke, based on lower-body features.

(f) Duke ▶ Mkt, based on lower-body features.

Figure 4.3: Mutual information between pseudo labels and camera index evolution for the SSG [32] setting. Ground-truth ID comparison is displayed in dashed lines for both datasets.

also observe that **CANU**-SSG's labels are slightly more camera dependant that SSG+Adv, even if the difference is less clear than for MMT. We also note that the camera dependency of the clustering results is much closer to the target mutual information.

4.5.5 Impact of CANU on lost IDs

Lost IDs on CANU-MMT Since we train the target dataset using unsupervised techniques, we do not use the ground-truth labels in the target dataset during training. Instead, we make use of the pseudo labels provided by DBSCAN. DBSCAN discards the outliers i.e. features that are not closed to others. It is natural to wonder how many identities are "lost" at every iteration. We here visualize the number of lost ID (all those that are not present in a training epoch) after each clustering step. We plot the evolution of this number with the training epoch for MMT, MMT+Adv. and **CANU**-MMT on Duke ▶ Mkt in Fig. 4.4. The dual experiment, i.e. on Mkt ▶ Duke revealed that no ID was lost by any method. In Fig. 4.4, we first observe that the loss of person identities decreases with the clustering steps. It means that the feature extractor provides more and more precise features representing person identities. Secondly, the use of camera adversarial training can reduce the loss of person identities in the clustering algorithm, which reflects the benefit of camera adversarial networks to the clustering algorithm and thus to the unsupervised person re-ID task.

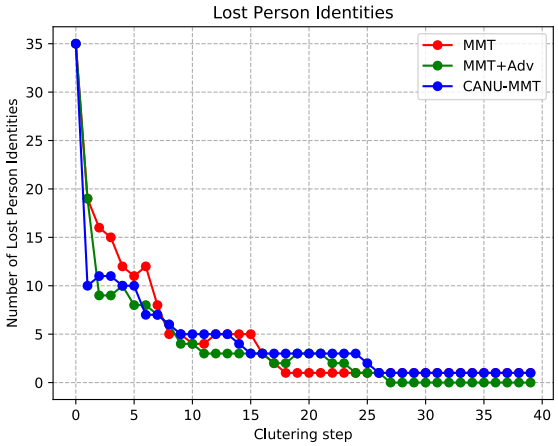
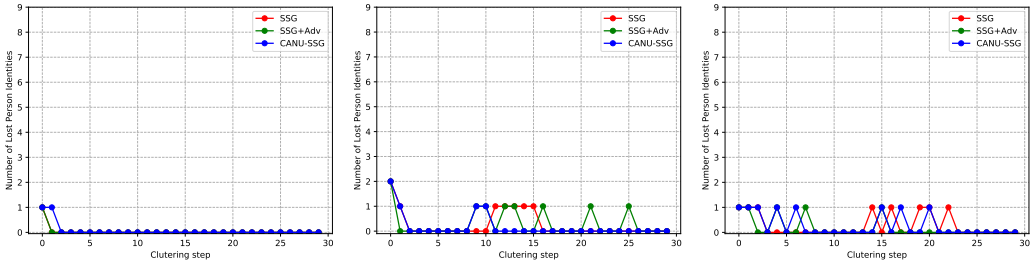


Figure 4.4: Evolution of the number of lost person IDs during training using MMT on Duke ▶ Mkt.



(a) # lost IDs based on full-body features. (b) # lost IDs based on upper-body features. (c) # lost IDs based on lower-body features.

Figure 4.5: Evolution of the number of lost person IDs during training using SSG [32] on Duke ▶ Mkt. In SSG, DBSCAN clusters full- (a), upper- (b) and lower-body (c) features independently.

Lost IDs on CANU-SSG As discussed previously, we do not use the ground-truth labels in the target data set during training: we make use of the pseudo labels provided by DBSCAN. Since SSG [32] uses three different types of features: full-, upper- and, lower-body features, we visualize in Fig. 4.5 the number of lost identities (the ground-truth identities that are not present in pseudo labels given by DBSCAN, i.e. they are considered as outliers.) respectively based on the clustering results from different features.

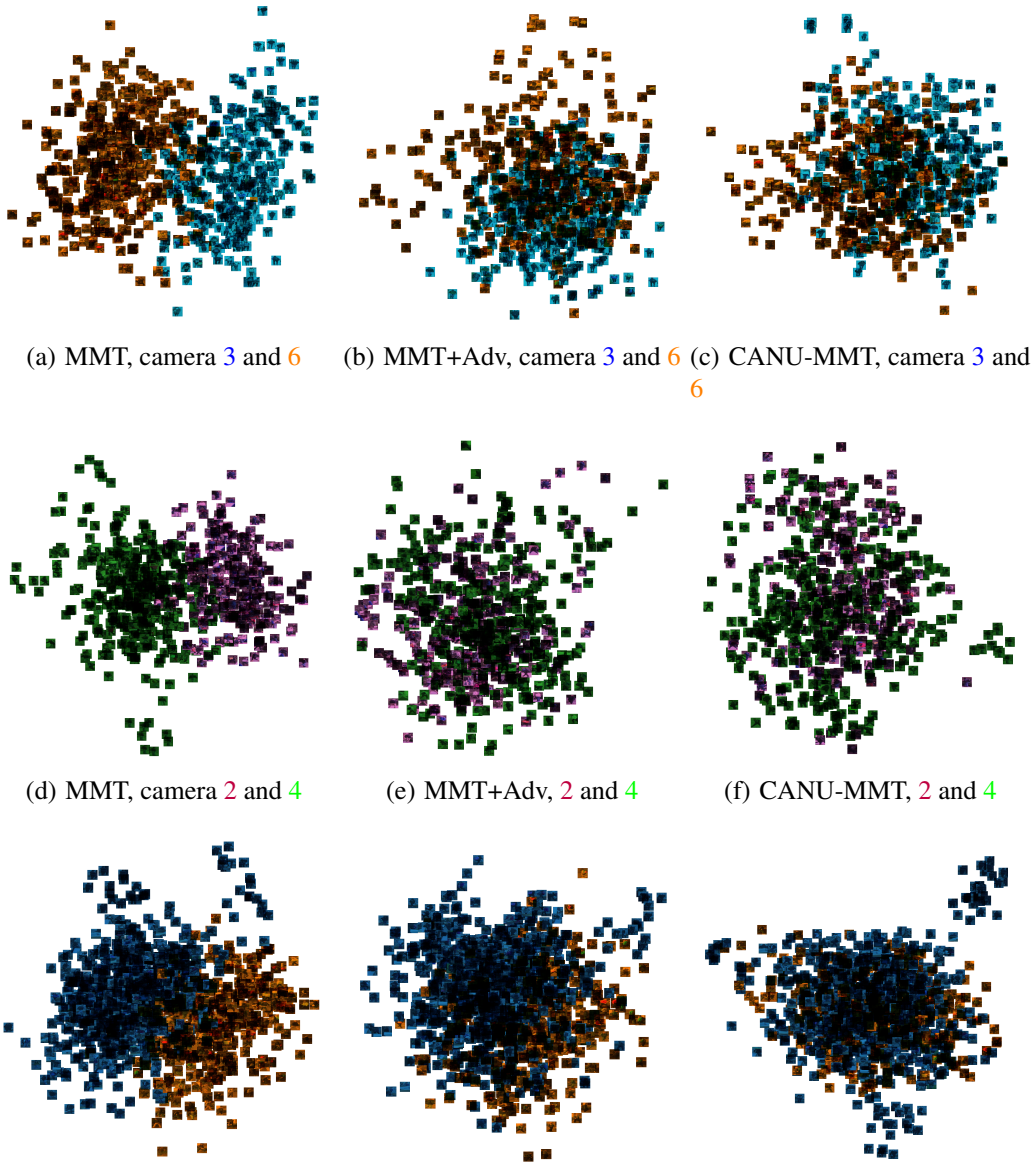
As in MMT, no IDs are lost in the Mkt ▶ Duke setting. For Duke ▶ Mkt, we observe that only a very few IDs are lost (≤ 2 IDs). Precisely, the number of lost person IDs from the clustering results on full-body features remains 0 except at the beginning of the first epoch (1 lost ID). Moreover, for upper-body features, All settings lose less than 2 IDs and remarkably, **CANU**-SSG loses no IDs during most of the training epochs and it has a lower loss compared to SSG+Adv and SSG. Finally, for lower-body features, at most 1 ID is lost during the training procedure. In summary, (1) very few IDs have been lost during on SSG using DBSCAN. (2)**CANU**-SSG has fewer losses compared to SSG+Adv and SSG. (e.g. Fig. 4.5 (b)).

4.5.6 EMBEDDING VISUALISATION

We use the re-ID features to compute a PCA projection of the embedding space for different pairs of cameras, and use the first 2 dimensions for visualization, in Figure 4.6. We use PCA to preserve the global structure, which is not guaranteed in other dimensionality reduction methods like t-SNE. We report the Mkt ► Duke setting for MMT, MMT+Adv, and **CANU**-MMT, using the train set of the target dataset. We observe that for all pairs of cameras, the embedding vectors distributions overlap more for both adversarial strategies, compared to the original implementation of MMT. It shows that our adversarial strategies achieve distributions matching across cameras more reliably than MMT alone.

4.6 CONCLUSION

In this chapter, we demonstrate the benefit of unifying adversarial learning with current unsupervised clustering-based person re-identification methods. We propose to condition the adversarial learning with the cluster centroids, being these representations independent of the number of clusters and invariant to cluster index permutations. The proposed strategy boosts existing clustering-based unsupervised person re-ID baselines and sets the new state-of-the-art performance in four different unsupervised person re-ID experimental settings.



(a) MMT, camera 3 and 6 (b) MMT+Adv, camera 3 and 6 (c) CANU-MMT, camera 3 and 6

(d) MMT, camera 2 and 4 (e) MMT+Adv, 2 and 4 (f) CANU-MMT, 2 and 4

(g) MMT, camera 0 and 6 (h) MMT+Adv, camera 0 and 6 (i) CANU-MMT, camera 0 and 6

Figure 4.6: PCA visualization of the embedding for Mkt ► Duke setting. Best viewed in color.

CHAPTER 5

DOMAIN ADAPTIVE MOT

5.1 INTRODUCTION

We propose to explore the domain adaptation framework leveraged in previous chapters to solve domain adaptation for Multiple Object Tracking, instead of re-ID alone. While unsupervised domain adaptation [33, 97, 13] and multiple object tracking [117, 127, 10, 106] have been widely explored respectively in the literature, up to our knowledge, no strategies have been proposed to deal with the domain shift problem in Multiple Object Tracking (MOT) in an unsupervised manner. We explore unsupervised domain adaptation for MOT: a labeled *source* MOT dataset and an unlabeled *target* MOT dataset are leveraged to optimize tracking performance on the latter. Tackling this specific problem is particularly relevant to real-world applications: collecting tracking annotations to adapt a tracker to new settings can be costly and troublesome, or using the same tracker for a long period of time might see the target domain shift significantly from the initial training domain, due to changing weather or illumination conditions.

Recent Multiple Object Tracking models rely on jointly solving the detection and the tracking task [117, 127, 10, 106]: recent increase in performance can partly be attributed to the detection recall improvement leveraged by such approaches. Importantly, previous works in the field of Domain Adaptative Object Detection [20, 92, 17, 104, 130, 81] highligthed the drop in detection performance in the case of a significant domain shift (due to changing weather or illumination conditions). As shown by Van Gool et al [20], and similarly to Unsupervised Domain Adaptation (UDA) approches, this problem can be adressed through Adversarial Domain Adaptive Detection, which casts it as a distribution alignment problem and solves it using an adversarial approach inspired by Ganin et al [33].

In parallel, most modern trackers use re-identification models to limit track fragmentation [127, 105] or exploit them as a cue, along with spatial information, to associate detections across time [10, 117]. As shown in previous chapters, person re-identification models lack generalization ability, especially when handling large visual variations (illuminations or background changes). This problem is generally being tackled in the

unsupervised person re-identification (URID) field, which introduced the *clustering* and *finetuning* framework as a possible solution.

We investigate for the first time the problem of unsupervised domain adaptation for MOT, and address it through joint adaptation of the re-identification and detection embeddings, casting the *clustering* and *finetuning* framework into a *tracking* and *finetuning* strategy, while leveraging adversarial strategies to perform adaptation.

Our contributions are the following:

- We demonstrate the domain-shift problem on standard MOT datasets via its impact in tracking performance when transferring a tracker from one dataset to another.
- We introduce a *tracking* & *finetuning* (T&F) framework to jointly adapt the re-identification and the detection embeddings: we alternate between generating pseudo-tracks on the target dataset using the current weights of tracker, and finetune it by leveraging adversarial strategies and pseudo-tracks to adapt the detection branch.
- We assess the MOT performance of our proposed T&F adaptation framework on two unsupervised domain adaptation setting using two widely used MOT datasets, namely MOT17 $\rightarrow$ MOT20 and MOT20 $\rightarrow$ MOT17, and demonstrate the superiority of our approach compared to direct transfer and simpler adaptation strategies.

5.2 Related Work

5.2.1 Multiple Object Tracking

Initial works in the MOT community focuses on optimally associating detections provided by standard detectors through a probabilistic framework [9, 88, 6]. Deep Learning methods initially propose to learn the association task by using Reccurent Network [77], and by modeling appearance and interaction to better capture track dynamics [128]. Motivated by the performance improvement in detection models, Tracktor [10] proposes to adapt the FRCCN architecture to perform MOT, and casts it as a regression task. Incidentally, it allows to refine public detections, and thus benefits from FRCCN's improved detection ability. Another familly of trackers [117, 127] builds upon the improved detection performance of CenterNet [128], fetching detections through a dedicated branch trained to perform heatmap regression. The detection-track association can be learned [127], or performed through a dedicated re-ID branch combined with a probabilistic framework [117]. More recently, DETR [109] proposes to use a Transformer [98] architecture to perform detection and paves the road for a new familly of transformer-based trackers [95, 75, 105].

Some methods use an external person re-ID network [10], or a re-ID branch as appearance descriptor [106, 105] to boost tracking performance. This motivates [54] to explore unsupervised MOT by finetuning a person re-ID network by using pseudo-label generated through a pretrained tracker.

5.2.2 Domain Adaptive Object Detection

Deep learning methods draw regained interest and rapid improvement in the object detection field [87, 86, 38], especially thanks to the Region of Interest proposal framework [37] combined with RoI pooling [38], as well as the use of RPN network to generate RoIs [87]. Adversarial development in the UDA community [33], combined with a need to develop methods leveraging unlabeled data to improve real-world deployment led to the developpement of Domain Adaptive Object Detection. Van Gool et al [20] first proposes to use adversarial domain adaptation for object detection, applied to a FRCNN [87] detector, performing distribution alignment at both RPN/image-level and regressor/instance-level. It paves the road for a Domain Adaptation Faster-RCNN family [92, 17, 104, 130, 81]: [92] performs differentiated alignment at local and global scale, [17] introduces a Hierarchical Calibration Network to balance feature transferability and discriminability, [104] trains an additional image-level categorical classifier to improve categorical consistency when performing instance-level alignement while [130] use clustering methods at the RPN level to adjust region-level alignment. Other line of works take advantage of CycleGAN's [129] ability to transfer image style to create and finetune the detector on a synthetic dataset [50, 46], or cast the task as a robust learning problem [55].

5.2.3 Unsupervised Person re-ID

Unsupervised person re-identification (URID) takes advantage of recent achievements in supervised person re-ID models without requiring the costly labeling process of the target dataset. Recent works start from a pretrained model on the source dataset and are based on the *clustering* and *finetuning* framework [28, 70, 32, 35, 115]. It alternates between a clustering step generating noisy pseudo-labels, and a fine-tuning step adapting the network to the target dataset, progressively refining pseudo-labels. A lot of focus has been in improving the quality of the pseudo-labels [28, 70, 32, 35].

A different approach is directly inspired by Unsupervised Domain Adaptation (UDA) [25, 124, 16, 84, 93, 125] and uses both the source and target datasets during adaptation. It aims at matching source and target distributions while keeping features discriminative ability by leveraging source ground truth ID labels. A first exemple learns to map source's detections to target's style detections, and train a re-ID model in a supervised setting using only those transferred detections [25], or in combination with the original target detections [124]. Alternatively, some exploit standard UDA strategies based on adversarial learning to match the source and target distributions [34, 84]. However, Negative Transfer [23] induced by the strict ID separation between source and target datasets limits the ability of such strategies.

An alternative approach exploits spatio-temporal consistency to re-construct tracklets from the dataset and use them to finetune the model [73, 61]. We propose to get inspiration from this strategy and combine it with *Clustering* and *Finetuning* strategies to propose a *Tracking* and *Finetuning* framework.

5.3 METHODOLOGY

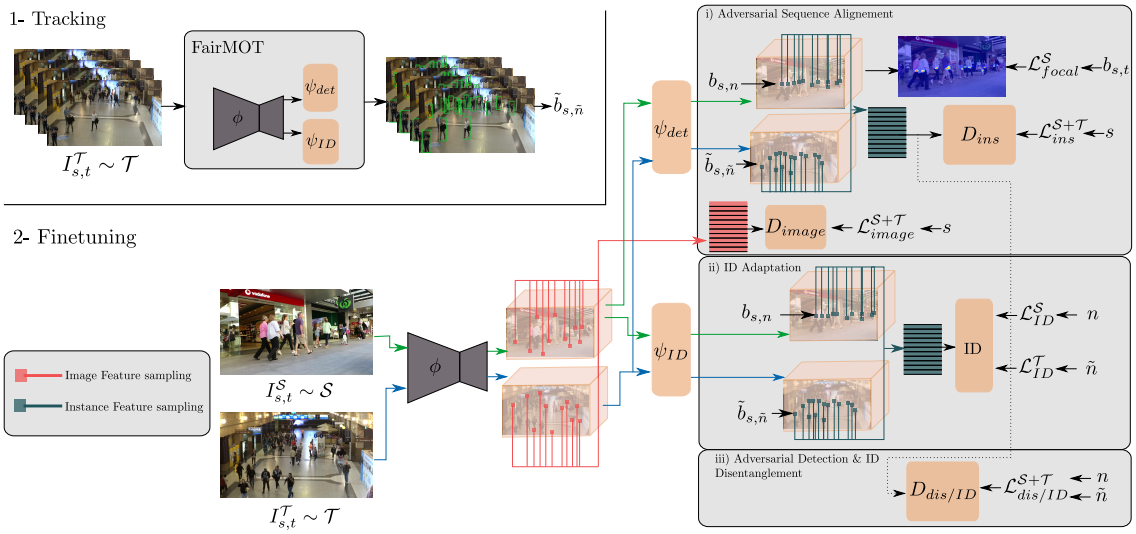


Figure 5.1: Overview of our proposed *Tracking* and *Finetuning* framework: the first step generates pseudo-tracks $\tilde{p}_{s,t}$ on all target sequences s using pretrained tracker weights. The second step finetunes the tracker using source ground truth annotation $p_{s,t}$ and pseudo-tracks $\tilde{b}_{s,t}$ to perform i) Adversarial Sequence Alignment to adapt the detection branch, ii) finetune the ID branch, and iii) enforce Adversarial Detection & ID Disentanglement: i) leverages annotations to extract detection and image features to perform adversarial alignment using D_{image} and D_{ins}, ii) uses a standard classification loss to adapt the ID classifier, and iii) uses a discriminator trained $D_{dis/ID}$ to extract ID information from detection instances and improve disentanglement. GRL are not displayed and considered as first layer of D_*.

We propose to address the problem of Domain Adaptation MOT with a *tracking* and *finetuning* framework, while leveraging adversarial strategies to align source and target datasets on the detection embedding. We first introduce the notations and the tracking framework we choose to build upon.

5.3.1 NOTATIONS AND TRACKER OVERVIEW

Let $\mathcal{S}$ be a source tracking dataset constituted by $M^{\mathcal{S}}$ sequences filled with images $\boldsymbol{I}^{\mathcal{S}}_{s,t}$, and annotated with $N^{\mathcal{S}}$ individual tracks denoted $\{b^{\mathcal{S}}_{s,n} = \{(x,y,w,h)^{\mathcal{S}}_{s,n,t}\}^{T_{s,n}}_{t=t_{0,n,s}}\}^{N^{\mathcal{S}}}_{n=1}$, x, y, w, h describing bounding box center position and size, s the sequence identifier of the track, and t the time frame at which the bounding box is extracted. Similarly, we note $\mathcal{T}$ a target tracking unannotated dataset, constituted by $M^{\mathcal{T}}$ sequences filled with images $\boldsymbol{I}^{\mathcal{T}}_{s,t}$. In order to smoothly merge a URID framework with a Domain Adaptative Object Detector, we need to build upon a tracker which jointly performs person re-identification and person detection. Furthermore, we focus on online trackers, since our main motivation is to investigate MOT in real world applications.

FairMOT [117] is a strong baseline illustrated in Figure 5.1 matching these requirements: a backbone ϕ first takes as input the current image (from source or target) and

produces a feature heatmap thanks to a shared embedding. It is then fed to a detection branch ψ_{det} and a re-ID branch ψ_{ID}: the detection branch is trained similarly to [128] with the focal loss [69] to regress heatmaps containing tracks center positions. The re-ID branch extracts ID features at $(x, y)_{s,t}$ positions from a feature heatmap and feed them to an ID classifier trained with a standard cross-entropy loss supervised by the ID labels n. Extra-branches such as Bounding Box size used in the original implementation are not detailed for simplicity, and are trained as in [117] if not mentioned otherwise. Similarly to state-of-the-art URID framework [23, 32, 35], we preliminarily pretrain FairMOT on the source dataset using the original training strategy [117]:

$$\mathcal{L}^{\mathcal{S}}_{pre}(\phi, \psi_{det}, \psi_{ID}) = \mathcal{L}^{\mathcal{S}}_{focal}(\phi, \psi_{det}) + \mathcal{L}^{\mathcal{S}}_{CE}(\phi, \psi_{ID}) \tag{5.1}$$

where $\mathcal{L}^{\mathcal{S}}_{focal}$ stands for the focal loss, and $\mathcal{L}^{\mathcal{S}}_{CE}$ for the categorical cross-entropy loss. Refer to [117] for training details. At inference and at each time step, the model uses spatial proximity and appearance similarity to match detections at t with known tracks at t-1, and Kalman filters to update track positions.

5.3.2 Tracking and Finetuning

We draw inspiration from the *clustering* and *finetuning* framework developed in the URID community, and adapt it to the MOT setting. The original formulation considers the embedding proximity of $\mathcal{T}$'s detections to perform clustering with a clustering algorithm and generates pseudo-ID labels that can be exploited to finetune the re-ID model. One of the limitations of this approach is that the clustering process is noisy due to the fact that the ID embedding has a limited knowledge of the target domain. Multiple iterations must therefore be performed in order to alleviate the mislabelisation problem, as seen in Chapter 4.

We propose to exploit spatial and temporal information in order to produce pseudo-ID labels, and replace the clustering step by a tracking step. It both takes into consideration the appearance similarity (by exploiting re-ID embedding when generating pseudo-tracks), and spatio-temporal information, which improves the quality of the pseudo-ID labels by forcing them to be temporally and spatially consistent. Namely, the tracking step generates the pseudo-tracks on the whole target dataset $\{\tilde{b}^{\mathcal{T}}_{s,\tilde{n}} = \{(\tilde{x}, \tilde{y}, \tilde{w}, \tilde{h})^{\mathcal{T}}_{s,\tilde{n},t}\}^{T_{s,\tilde{n}}}_{t=t_{0,\tilde{n}}}\}^{\tilde{N}^{\mathcal{T}}}_{\tilde{n}=1}$ using the source pretrained tracker and the inference strategy described in [117].

To finetune the re-ID branch, we replace the original ID classifier by an ID classifier of size $N^{\mathcal{S}} + \tilde{N}^{\mathcal{T}}$ and define the associated re-ID loss as the mean source ID cross-entropy

and target cross-entropy:

$$\mathcal{L}_{ID}^{\mathcal{S}+\mathcal{T}}(\phi,\psi_{ID}) = \frac{1}{2}(\mathcal{L}_{ID}^{\mathcal{S}}(\phi,\psi_{ID}) + \mathcal{L}_{ID}^{\mathcal{T}}(\phi,\psi_{ID})) \tag{5.2}$$

$$\text{with } \mathcal{L}_{ID}^{\mathcal{M}}(\phi,\psi_{ID}) = -\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{M}}\sim\mathcal{M}}\left\{\sum_{\tilde{n}\in\boldsymbol{I}_{s,t}^{\mathcal{M}}} \log\left\langle(\psi_{ID}(\phi(\boldsymbol{I}_{s,t}^{\mathcal{M}}))_{\lfloor\tilde{b}_{s,\tilde{n},t}\rfloor}), \tilde{\boldsymbol{n}}^{\mathcal{M}}\right\rangle\right\}$$

where $\mathcal{M} \in \{\mathcal{S},\mathcal{T}\}$, and $h_{\lfloor x\rfloor}$ refer to the feature vector at position x in the feature heatmap h.

5.3.3 Adversarial Sequence Alignment

We adapt the adversarial framework proposed in [20] to finetune the detection branch, in order to align source and target embedding distribution. Two domain discriminators are introduced, an *Instance-level* discriminator D_{ins} and an *Image-level* discriminator D_{image}. The former is fed with detection features extracted from the detection branch ψ_{det} and the latter with image features extracted from the shared embedding ϕ. They are equipped with a Gradient Reversal Layer [33] (GRL) and were originally being trained to recover domain information from input features.

Sequence Alignment As demonstrated in chapter 3, using a multi-class discriminator to perform distribution matching lead to better alignment than standard two-class discriminators. This is especially the case when the training dataset is constituted of images extracted from a small number of distinguishible visual domains (cameras or sequence ID). The underlying assumption is that breaking down intra-domain embedding structure facilitates inter-domain distribution alignment[84]. We note that MOT datasets, contrary to detection datasets, are constituted of recordings of a dozen of video sequences, where each sequence has distinguishable and temporally consistent background, illumination conditions and camera viewpoints. Also, we propose to perform *Instance-level* and *Image-level* alignment to align embedding distributions on a sequence basis. From an implementation perspective, we change discriminators final layers by a classification layer of size $M^{\mathcal{S}} + M^{\mathcal{T}}$ activated with a softmax function, and use a categorical cross-entropy loss to train D_{image} and D_{ins}.

Instance-level alignment The original approach is based on a FRCNN detector, and selects instance-level features thanks to locations provided by the RPN block. The downside of this strategy is that the RPN only provides rough estimates of the object position, and is prone to produce noisy detections. The *tracking* and *finetuning* framework provides temporally consistent pseudo-tracks annotation less noisy than detections. Thanks to the fully

convolutional nature of FairMOT's detection branch ψ_{det}, instance-level features are extracted using pseudo-tracks positions in the heatmap features located right before the final heatmap output of the detection branch. Once extracted, they are fed to the discriminator D_{ins}, trained using $\mathcal{L}_{ins}^{\mathcal{S}+\mathcal{T}}$ detailed in (5.3).

$$\begin{aligned}
&\mathcal{L}_{ins}^{\mathcal{S}+\mathcal{T}}(\phi,\psi_{det}, D_{ins}) = \\
&-\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{T}}\sim\mathcal{T}}\left\{\sum_{\tilde{n}\in\boldsymbol{I}_{s,t}^{\mathcal{T}}} \log\left\langle D_{ins}(\psi_{det}(\phi(\boldsymbol{I}_{s,t}^{\mathcal{T}}))_{\lfloor\tilde{b}_{s,\tilde{n},t}\rfloor}), \boldsymbol{s}\right\rangle\right\} \\
&-\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{S}}\sim\mathcal{S}}\left\{\sum_{n\in\boldsymbol{I}_{s,t}^{\mathcal{S}}} \log\left\langle D_{ins}(\psi_{det}(\phi(\boldsymbol{I}_{s,t}^{\mathcal{S}}))_{\lfloor b_{s,n,t}\rfloor}), \boldsymbol{s}\right\rangle\right\}
\end{aligned} \tag{5.3}$$

Image-level alignment We perform image-level alignment on the shared embedding produced by ϕ by sampling random position to extract image features from the feature heatmap. They are then fed to D_{image}, which is trained with (5.4).

$$\begin{aligned}
\mathcal{L}_{image}^{\mathcal{S}+\mathcal{T}}(\phi, D_{image}) = &-\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{T}}\sim\mathcal{T}, f^{\mathcal{T}}\sim\phi(\boldsymbol{I}_{s,t}^{\mathcal{T}})}\left\{\log\left\langle D_{image}(f^{\mathcal{T}}), \boldsymbol{s}\right\rangle\right\} \\
&-\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{S}}\sim\mathcal{S}, f^{\mathcal{S}}\sim\phi(\boldsymbol{I}_{s,t}^{\mathcal{S}})}\left\{\log\left\langle D_{image}(f^{\mathcal{S}}), \boldsymbol{s}\right\rangle\right\}
\end{aligned} \tag{5.4}$$

In order to maintain the detection ability of the branch on the source dataset while matching embedding distribution, we further supervise the detection branch with the focal loss $\mathcal{L}_{focal}^{\mathcal{S}}(\phi, \psi_{det})$ introduced in loss (5.1), evaluated on the data batch sampled from the source.

Overall, the detection part of the model is trained using the loss (5.5):

$$\begin{aligned}
\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}(\phi, \psi_{det}, D_{ins}, D_{image}) = &-\mathcal{L}_{ins}^{\mathcal{S}+\mathcal{T}}(\phi, \psi_{det}, D_{ins}) \\
&-\mathcal{L}_{image}^{\mathcal{S}+\mathcal{T}}(\phi, D_{image})+\mathcal{L}_{focal}^{\mathcal{S}}(\phi, \psi_{det})
\end{aligned} \tag{5.5}$$

5.3.4 Adversarial learning for disentanglement

So far we considered adapting the detection and re-ID branches separately, that is performing sequence alignment on the detection branch using losses (5.3) and (5.4), and finetuning the re-ID branch using the IDs of pseudo-tracks and the loss (5.2). We propose to also consider the influence of ID information on the detection branch when performing distribution alignment. Since the backbone is shared to perform both tasks, it is reasonable to assume that without strong supervision from target annotations, detection and ID

information are not properly disentangled in the detection branch. Also, because ID information is domain specific, improving disentanglement should improve the generalization ability of the detection branch, and thus improve its transfer ability.

We propose to leverage ID or pseudo-ID associated to bounding box annotation (or pseudo-annotation) in an adversarial setting: similarly to *Instance level* alignment, feature vectors are extracted from the feature heatmap produced by ψ_{det} at position given by track/pseudo-tracks annotations, and fed to an ID-discriminator $D_{dis/ID}$. Equipped with a GRL, $D_{dis/ID}$ is adversarially trained to recover ID information from the detection features. By doing so, ID and detection embedding disentanglement is enforced, forcing detection features to be ID invariant.

$$\begin{aligned}
&\mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}(\phi,\psi_{det}, D_{dis/ID}) = \\
&-\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{T}}\sim\mathcal{T}}\left\{\sum_{\tilde{n}\in\boldsymbol{I}_{s,t}^{\mathcal{T}}} \log\left\langle D_{dis/ID}(\psi_{det}(\phi(\boldsymbol{I}_{s,t}^{\mathcal{T}}))_{\lfloor\tilde{b}_{s,\tilde{n},t}\rfloor}), \tilde{\boldsymbol{n}}^{\mathcal{T}}\right\rangle\right\} \\
&-\mathbb{E}_{\boldsymbol{I}_{s,t}^{\mathcal{S}}\sim\mathcal{S}}\left\{\sum_{\tilde{n}\in\boldsymbol{I}_{s,t}^{\mathcal{S}}} \log\left\langle D_{dis/ID}(\psi_{det}(\phi(\boldsymbol{I}_{s,t}^{\mathcal{S}}))_{\lfloor b_{s,n,t}\rfloor}), \boldsymbol{n}^{\mathcal{S}}\right\rangle\right\}
\end{aligned} \tag{5.6}$$

5.3.5 OVERALL TRAINING

As illustrated in Figure 5.1 we solve the following adversarial problem:

$$\begin{aligned}
\min_{\psi,\psi_{ID},\psi_{det}} \max_{D_{ins},D_{image},D_{dis/ID}} &\mathcal{L}_{ID}^{\mathcal{S}+\mathcal{T}}(\phi,\psi_{ID})+ \\
&\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}(\phi,\psi_{det},D_{ins},D_{image}) - \mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}(\phi,\psi_{det},D_{dis/ID})
\end{aligned} \tag{5.7}$$

5.4 EXPERIMENTAL RESULTS

5.4.1 IMPLEMENTATION DETAILS

The original FairMOT implementation is used to conduct experiments, along with the original training strategy. The original DLA-34 [128] variant is used as backbone, and pretraining strategy is kept exactly as described in [117]. For target adaptation, we use the Adam optimizer [56], with a learning rate of 10^{-4} during 20 epochs, with a batch size of 8 (4 from source, 4 from target). Standard augmentation techniques are used, such as random rotation, scaling and color jittering, on both source and target input images. To perform fair evaluation across adaptation settings and training strategies, different detection threshold are considered and evaluated on the validation set, and best performing MOTA is reported for each strategy and adaptation setting. We train the model on two Titan X GPUs for about 10 hours.

5.4.2 EVALUATION PROTOCOL

Datasets and metrics We use standard MOT datasets MOT17 [76] and MOT20 [24] along with their official splits. Two Domain Adaptation settings are considered: MOT17 $\rightarrow$ MOT20 and MOT20 $\rightarrow$ MOT17, where MOT17 is the annotated source and MOT20 the unannotated target, and vice-versa. Because ground truth annotation is not available on both test sets, the ablation study is performed on MOT17 and MOT20's training sets similarly to [117]: only the first half of both training sets are then used for training, the validation performance being computed on target's training second half. The official test server is used to evaluate MOT performance on the test set.

The MOT17 test set contains 2,355 trajectories, scattered across 7 sequences and a total of 17,757 frames, while MOT20 test set contains 1,501 trajectories, across 4 sequences of 4,479 frames. Noticeably, the trajectory density (ie the average number of tracks present in each frame) is 31.8 for MOT17 and 170.9 for MOT20: in addition to the large background and illumination condition changes between MOT17 and MOT20, the model has also to consider a drastic detection distribution shift, making this setting particularly interesting to investigate Domain Adaptation for MOT.

Standard MOT metrics are reported to evaluate the performance of our approach: MOTA [11] (Multiple Object Tracking Accuracy) assesses how well proposed tracks match GT annotation by taking into account FPs (False Positives), FNs (False Negatives) and IDs (identity switches). MOTP assesses how far from matched GT tracks proposal are. IDF1, IDR and IDP (ID F1, Recall and Precision), which assess how well the tracker performs in terms of identification ability [89] are also reported. We perform the experiments in the private detection setting as is FairMOT.

Table 5.1: Domain Adaptation MOT test results for MOT17 → MOT20 setting: the official test set of the MOT20 dataset is used. Direct Transfer refers to MOT performance of FairMOT trained on the full training set MOT17. Best results are in **bold**.

MOT17→MOT20	IDF1	Rcll	Prcn	IDs	MOTA	MOTP
Direct Transfer [117]	42.1	48.3	91.1	4,659	42.7	77.0
$\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}} + \mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}$	**59.6**	71.2	78.5	3,998	**51.0**	75.6

5.4.3 RESULTS AND DISCUSSION

Test set evaluation We report in Table 5.1 the performance of our approach under the MOT17 → MOT20 setting, using the official test set of MOT20, and the full MOT17 and MOT20 training sets. We compare with a direct transfer strategy, where the tracker is trained using the original training scheme on the whole MOT17 training set. We observe that domain adaptive MOT significantly improves target tracking performance, by +8.3% in MOTA and +17.5% in IDF1. We note that detection recall significantly improves with the domain adaptive approach by +22.9%, leading to a higher track recall and thus explaining the higher MOTA. The higher identification performance is partly explained by the lower number of identity switches, and is not surprising considering the poor generalization ability of pretrained re-ID models, as developed in chapter 3: using target's noisy pseudo-labels largely benefits the re-ID branch to adapt to the new domain.

Ablation Study We run adaptation experiments for both MOT17 → MOT20 and MOT17 → MOT20 in Table 5.2 and Table 5.3, using different training strategies to assess their role: we first replace Adversarial Sequence Alignement $\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}$ with Adversarial Domain Alignment $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$ as in [20], where the loss of the D_{ins} and D_{image} is replaced by a binary cross-entropy loss and supervised with domain labels, and train the model without ID-disentanglement; we then assess Adversarial Sequence Alignment $\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}$ without ID-disentanglement; and finally train the model with both Sequence Alignment and ID-disentanglement: $\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}} + \mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}$. For the MOT17→MOT20 setting, the pseudo-tracks are re-generated every 7 epochs, while for MOT20→MOT17 they are generated only once.

We note that $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$ already improves target performance compared to direct transfer, in MOTA by +3.8% and +3.2% respectively in MOT20 and MOT17, in IDF1 by +11.8% and +1.9%. The large gap in IDF1 improvement is due to the larger track density in MOT20, which yields a higher risk of ID switches and has a significant impact on IDF1. Similarly, $\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}$ improves tracking performance compared to $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$, by +3.7% and +1.4% in MOTA and +3.3% and +0.8% in IDF1. Finally, we note that the addition of ID disentanglement $\mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}$ consistently improves results in MOTA, by +0.5% and +0.8%, but slightly lowers IDF1 in MOT20 by -0.2%, while improving IDF1 for MOT17 by +0.7%. Compared to Direct Transfer, the proposed approach improves MOTA by +8.0% and +5.4%, IDF1 by 14.9% and 3.4%.

Table 5.2: Domain Adaptation MOT results for MOT17 → MOT20 setting: the second half of of the MOT20 dataset is used as validation set. Direct Transfer refers to MOT performance of FairMOT trained on MOT17. $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$, replacing Adversarial Sequence Alignement, refers to Adversarial Domain Alignment as in [20]. Best results are in **bold**.

MOT17→MOT20	IDF1	IDP	IDR	Rcll	Prcn	IDs	MOTA	MOTP
Direct Transfer [117]	51.3	63.9	42.8	61.2	91.3	6952	54.4	24.4
$\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$	63.1	71.9	56.2	68.4	87.6	3888	58.2	25.7
$\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}$	**66.4**	74.1	60.1	71.8	88.4	3303	61.9	25.4
$\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}} + \mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}$	66.2	75.3	59.0	70.5	90.1	3085	**62.4**	23.8

Table 5.3: Domain Adaptation MOT results for MOT20 → MOT17 setting: the second half of of the MOT17 dataset is used as validation set. Direct Transfer refers to MOT performance of FairMOT trained on MOT20. $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$, replacing Adversarial Sequence Alignement, refers to Adversarial Domain Alignment as in [20]. Best results are in **bold**.

MOT20→MOT17	IDF1	IDP	IDR	Rcll	Prcn	IDs	MOTA	MOTP
Direct Transfer [117]	64.6	81.1	53.7	60.1	90.7	188	53.6	22.4
$\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$	66.5	84.2	54.9	61.2	93.9	212	56.8	25.6
$\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}}$	67.3	83.0	56.5	63.3	93.0	209	58.2	25.9
$\mathcal{L}_{seq}^{\mathcal{S}+\mathcal{T}} + \mathcal{L}_{dis/ID}^{\mathcal{S}+\mathcal{T}}$	**68.0**	87.8	55.4	61.2	97.0	187	**59.0**	25.1

Pseudo-tracks generation analysis We further study the impact of the frequency of pseudo-tracks generation in Table 5.4. Interestingly, we note that the MOT performance does not behave similarly in both settings: in MOT17→MOT20, the more frequent the pseudo-track generation is, the better MOT performance is, suggesting that gradually improving the pseudo-tracks helps the model to converge to a better solution. On the contrary, in MOT20→MOT17, the performance only gets worse as pseudo-track generation gets more frequent. We explain this by the fact that MOT20 has a track distribution significantly different compared to MOT17. Per sequence track density for MOT20 goes from 70 to 205 with an average of 170.9, when for MOT17, track density goes from 9.6 to 69.8 with an average of 31.8. Also, aligning the distribution of the detection branch encourages the tracker to produce as many detections on MOT17 as in MOT20, and thus produce FPs, as shown by the diminishing precision, which is detrimental to pseudo-track generation. This is not the case in MOT17 → MOT20, since matching track density on MOT20 will only lead to FNs, which do not affect significantly our instance-level alignment strategy and thus the adaptation performance. We argue that the more realistic scenario is MOT17→MOT20, since a real-life source dataset should have lower-density tracks (annotation being less work-intensive, and most existing MOT dataset have low-density tracks) compared to the target dataset.

Table 5.4: Domain Adaptation MOT results for MOT20 → MOT17 and MOT17 → MOT20 settings: the second half of of the target dataset is used as validation set. We change the number of iterations of the *tracking* and *finetuning*, ie the number of time η the pseudo-tracks are regenerated during adaptation. The overall number of epochs remains fixed to 20 for all runs. Best results are in **bold**.

MOT17→MOT20	IDF1	IDP	IDR	Rcll	Prcn	IDs	MOTA	MOTP
$\eta = 1$	62.5	70.6	56.1	70.0	88.1	4316	60.0	22.7
$\eta = 2$	65.8	73.7	59.5	71.5	88.5	3602	61.7	23.1
$\eta = 3$	**66.2**	75.3	59.0	70.5	90.1	3085	**62.4**	23.8
MOT20→MOT17								
$\eta = 1$	**68.0**	87.8	55.4	61.2	97.0	187	**59.0**	25.1
$\eta = 2$	65.9	79.3	56.3	63.3	89.1	195	55.2	31.4
$\eta = 3$	62.4	76.0	52.9	62.0	89.1	263	54.0	32.5

5.5 CONCLUSION

In this chapter, we investigate unsupervised domain adaptive MOT, and propose the *tracking* and *finetuning* framework to address it. We combine this approach with an adversarial strategy to perform domain adaptation via sequence distribution alignment on the detection branch. The adversarial framework is further leveraged to perform ID and detection disentanglement. We quantitatively demonstrate the efficiency of the proposed approach on two standard MOT datasets and adaptation settings, namely MOT17 → MOT20 and MOT20 → MOT17, conduct an extensive ablation study to assess the impact of each loss, and compare our proposal with simpler adaptation strategies.

CHAPTER 6

CONCLUSION

6.1 SUMMARY

This Thesis investigates different frameworks for MOT domain adaptability. First, a generative probabilistic model combined with an online appearance learning scheme is proposed to perform MOT in robotic settings: the generative framework relaxes the domain dependency problem that standard discriminative MOT systems might endure, while the learned appearance model allows the system to adapt the tracker to the target domain on the fly, while demonstrating better identification performance than standard appearance model based on a pretrained fixed representation. Because state of the art MOT frameworks make use of person re-ID models to improve identity consistency, domain adaptability for re-ID models, namely Unsupervised Person re-ID, is then investigated. Inspired by domain-adversarial strategies developed for domain adaptation in classification tasks, a camera-adversarial strategy is devised, and demonstrates higher re-ID performance on standard target re-ID datasets than standard domain-wise adversarial approaches, and competitive performance with respect to state-of-the-art strategies. It further demonstrates the ability of the adversarial framework to perform distribution matching for more than two distributions. Then, the camera-adversarial strategy is investigated in a novel unsupervised re-ID framework: *clustering* and *finetuning*, where only the target domain is used during training. The camera-adversarial strategy reduces pseudo-labels mislabeling during the clustering step, and thus improves ID discriminability and target re-ID performance. The impact of negative transfer in such strategies is also demonstrated, that is the non-uniform ID distribution over camera domains, and a conditional adversarial approach to perform ID-conditioned camera distribution matching is proposed. Applied on two state-of-the-art learning strategies, it demonstrates better target re-ID performance against both their original performance and vanilla camera adversarial strategy, yielding new state-of-the-art performance in unsupervised re-ID. Finally, the adversarial framework is investigated to perform domain adaptation in MOT: the impact of the domain-shift problem in MOT on standard datasets is demonstrated, and addressed by leveraging adversarial methods inspired by domain adaptive object detection and a *tracking* and *finetuning* training scheme is proposed. An adversarial ID-detection disentanglement strategy

is also proposed to alleviate ID related features learning during detection branch finetuning. The superiority of this strategy compared to simpler adaptation scheme is experimentally demonstrated on two standard MOT datasets, and an extensive ablation study is performed to assess the role of each losses.

6.2 FUTURE RESEARCH DIRECTIONS

From this thesis, several lines of future work could be explored:

- Chapter 2: The probabilistic model used to derive the presented MOT algorithm can easily be extended for Audio-Visual tracking [8]. This setting has the advantage that audio information is not limited by the field of view of the camera and thus can track people all around the robot, and processing audio information is particularly interesting to better understand conversations. A possible line of future work could therefore be to implement a voice model similar to the appearance model developed in chapter 2, and trained online with self-annotated recordings. Further, recent works in the speaker recognition literature [80] demonstrated that visual information can be leveraged when processing audio-recordings to improve accuracy: a complementary approach could investigate how to jointly model identity using both visual and audio cues. Modeling those information together would be particularly interesting when performing cross-domain re-ID (a previously seen person start speaking, or a previously speaking person become visible).
- Chapter 3: Investigating the impact of camera-adversarial strategies highlighted the camera-dependency of standard person re-ID model, even when trained with ID supervision (i.e. in standard supervised person re-ID). However, preliminary experiments did not show improvement in such framework: even though camera embedding-invariance is enforced, it does not translate in improved re-ID performance. This is counter-intuitive, and remains to be investigated.
- Chapter 4: The adversarial approach devised to guide the unsupervised learning of person re-ID models could also be explored in the more general context of Unsupervised Learning. Especially, adversarial approaches could be leveraged in order to reduce embedding variability when heavy data-augmentation techniques are applied to the training data. Similarly, self-supervision [26] could be explored through the lens of adversarial proxy tasks: instead of solving tasks close to the aimed objective with labels generated from the training data, one could additionally solve tasks designed to enforce model invariance, exploiting labels also generated directly from training data.
- Chapter 5: The Domain Adaptive framework devised in this chapter could easily be applied to any other tracker: also a possible future work could explore the impact this strategy on trackers taking advantage of the higher detection ability leveraged by the transformer architecture [98, 109], like Transcenter[105]. Also, as seen in chapter 4, a wide variety of unsupervised strategy can be leveraged to perform unsupervised

person re-ID: an alternative idea could investigate how to integrate methods like MMT [35] or SSG [32] within the *clustering* and *finetuning* framework to improve pseudo-ID consistency and thus the ID-branch discriminative ability.

- More general questions common to all the chapters of this thesis could also be investigated: first, the computational aspect of domain adaptation is often overlooked: most of the presented strategies are applied during training, while the backbone model remains untouched. In order to enable robotic application, significant architecture simplification should be performed to reduce computational cost[45], and such changes could also benefit to adaptation performance since reducing modeling ability might benefit generalization ability. Second, continuous adversarial adaptation could be investigated to perform adaptation on the fly, for both detection and identification modules, as done in chapter 2. In this setting, the temporal aspect of adaptation should be taken into account, along with significant training simplification to allow for real-time adaptation. Finally, robot spatio-temporal information, leveraged by its ability to navigate its environment could also be exploited to further guide visual adaptation in long-term tracking settings, where drastic illumination conditions and background changes severely reduce the robot ability to keep ID switches low.

6.3 PUBLICATIONS AND SUBMISSIONS

Here is the list of papers that have been published or submitted during my PhD.

- [2] Xavier Alameda-Pineda, Soraya Arias, Yutong Ban, **Guillaume Delorme** and Laurent Girin, and Radu Horaud, Xiaofei Li, Bastien Morgue, and Guillaume Sarrazin, Audio-Visual Variational Fusion for Multi-Person Tracking with Robots, In *Proceedings of the 27th ACM International Conference on Multimedia*
- [22] **Guillaume Delorme**, Yutong Ban, Guillaume Sarrazin, Xavier Alameda-Pineda, ODANet: Online Deep Appearance Network for Identity-Consistent Multi-Person Tracking, In *2020 - 25th International Conference on Pattern Recognition (ICPR) / Workshops*
- [23] **Guillaume Delorme**, Yihong Xu, Stéphane Lathuiliére, Radu Horaud, Xavier Alameda-Pineda, CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification, In *2020 - 25th International Conference on Pattern Recognition (ICPR)*
- [127] Yihong Xu, Yutong Ban, **Guillaume Delorme**, Chuang Gan, Daniela Rus, Xavier Alameda-Pineda, TransCenter: Transformers with Dense Queries for Multiple-Object Tracking, submitted to *IEEE 2021 International Conference on Computer Vision (ICCV)*

CHAPTER A

APPENDIX: ADVERSARIAL LEARNING FOR MULTI-DOMAIN CLASSIFIER

In this appendix, we show that the adversarial formulation can be interpreted in terms of Jenson-Shannon divergence minimization.

The goal is to prove that optimizing the adversarial loss is equivalent to minimizing the Jensen- Shannon divergence (JSD) between feature distributions across cameras. It relates to the second term of the loss (3.4) in chapter 3. This equivalence has been proved in the original GAN formulation for binary discriminators [40]. We extend the derivation for an arbitrary number of distributions, and to the JSD generalized to multiple distributions.

Let ϕ be our feature extractor, and D_{cam} our camera classifier. Let's define $\mathcal{D}$ as our person Re-ID dataset. It is composed of 3-tuples $(x_n, c_n, p_n)_{n=1}^{N}$, which correspond to the images, the camera indexes, and the identities, respectively. We also denote with C the number of cameras. Our problem of camera feature distribution matching can be written as the adversarial optimization of

$$\min_{\phi} \max_{D_{cam}} E_{x,c\sim\mathcal{D}} \log D_{cam}(\phi(x))_c \tag{A.1}$$

During optimization, when the feature extractor ϕ is fixed, the camera classifier D_{cam} is optimized according to:

$$\max_{D_{cam}} E_{x,c\sim\mathcal{D}} \log D_{cam}(\phi(x))_c = \tag{A.2}$$
$$\int_x \sum_c p(\phi(x), c) \log(D_{cam}(\phi(x))_c) dx$$

For a given image input x, we seek an optimal camera classifier D^*_{cam} with posterior probability d^*, namely

$$d^* = \underset{d}{\operatorname{argmax}} \quad \sum_{c=1}^{C} p(\phi(x), c) \log(d) \tag{A.3}$$

with the constraints $d = (d_1, \ldots, d_C) \in \mathbb{R}^C$, $\forall c,\ 1 \leq c \leq C, 0 \leq d_c \leq 1, \sum_{c=1}^{C} d_c = 1$ that ensure that d describes a probability distribution. For the sake of notation simplicity, we denote $f(d) = \sum_c p(\phi(x), c) \log(d)$. We also define $\lambda = (\lambda_1, \ldots, \lambda_C)$ and $\mu = (\mu_1, \ldots, \mu_C)$.

The problem of (A.3) turns out to be the optimization of the generalized Lagrangian [39], which is written as:

$$\begin{aligned} &\max_{d} \min_{\gamma,(\mu,\lambda\geq 0)} \mathcal{J}(d, \lambda, \mu, \gamma) = \\ &f(d) + \sum_{c} \lambda_c d_c + \sum_{c} \mu_c (1 - d_c) + \gamma (\sum_{c} 1 - d_c) \\ &\nabla_{d_c} \mathcal{J}(d_c^*) = 0 \Leftrightarrow \frac{p(\phi(x), c)}{d_c^*} + \lambda_c - \mu_c - \gamma = 0 \end{aligned} \tag{A.4}$$

that leads to

$$\begin{aligned} d_c^* &= \frac{p(\phi(x), c)}{\mu_c - \lambda_c + \gamma} \\ &= \frac{p(\phi(x), c)(d_c^*)(1 - d_c^*)}{\mu_c(d_c^*)(1 - d_c^*) - \lambda_c(d_c^*)(1 - d_c^*) + \gamma(d_c^*)(d_c^* - 1)} \end{aligned}$$

Using the Karush-Kuhn-Tucker property [39], we have $\mu_c(1 - d_c^*) = 0$ and $\lambda_c d_c^* = 0$, thus

$$d_c^* = \frac{p(\phi(x), c)}{\gamma} \tag{A.5}$$

Due to the constraints of (A.3), when summing over the camera index c, we obtain:

$$\sum_{c} d_c^* = \sum_{c} \frac{p(\phi(x), c)}{\gamma} = 1 \tag{A.6}$$

where $\gamma = \sum_c p(\phi(x), c)$ is a normalization factor. Therefore the optimal posterior probability is equal to:

$$d_c^* = \frac{p(\phi(x), c)}{\sum_i p(\phi(x), c = i)} \tag{A.7}$$

After injecting the optimal discriminator D_{cam}^* from (A.7) above into (A.1), we obtain:

$$
\begin{aligned}
&\min_{\phi} E_{x,c\sim\mathcal{D}} \log \frac{p(\phi(x),c)}{\sum_i p(\phi(x),c=i)} \\
&= \int_x \sum_c p(\phi(x)|c)p(c) \log \frac{p(\phi(x)|c)p(c)}{\sum_i p(\phi(x),c=i)} dx \\
&= \underbrace{\sum_c p(c) \int_x p(\phi(x)|c) \log p(\phi(x)|c) dx}_{-\sum_c p(c) H(p(\phi(x|c)))} \\
&\quad + \underbrace{\sum_c p(c) \int_x p(\phi(x)|c) \log p(c) dx}_{-H(p(c))} \\
&\quad - \underbrace{\int_x \sum_c p(\phi(x)|c)p(c) \log \left(\sum_i p(\phi(x)|c=i)p(c=i) \right) dx}_{H(\sum_i p(c=i)p(\phi(x)|c=i))}
\end{aligned}
\tag{A.8}
$$

We recognize the definition of the JSD in the equation (A.8), and therefore conclude the proof, solving (A.1) is equivalent to solving:

$$
\min_{\phi} JSD_{p(c)}(p(\phi(x)|c=1), ..., p(\phi(x)|c=C)) \tag{A.9}
$$

where JSD is the generalized Jenson-Shannon divergence between the C distributions $p(\phi(x)|c)$. Note that each distribution $p(\phi(x)|c)$ is weighted by the prior distribution $p(c)$ in the computation of the generalized JSD.

To make sure that each camera distribution $p(\phi(x)|c)$ is considered equally, the prior distribution for each camera should be uniform $p(c) = \frac{1}{N_c}$. It highlights the importance of solving the unbalanced camera-classification problem in our problem. In practice, it is realized by sampling evenly across cameras.

LIST OF FIGURES

1.1 Datasets used in this thesis: (a) and (b) are MOT datasets, (c), (d) and (e) are re-ID datasets. 14

2.1 The robotic software architecture is composed of several nodes: an image is produced by the video driver, fed to the face detector which produces both face detections and appearance features, then transmitted to the tracker node (alongside motor position information). The tracking results are exploited by the robot control to move the robot's head exploiting the motors drivers. 26

2.2 Tracking qualitative results using CH and ODA-UP. Detections are displayed on the left panel (blue), and tracking results are available on the right panel (green) in 2 settings. 31

3.1 We propose an adversarial method to learn a feature extractor defining an embedding space where the different camera distribution match, and are not distinguisible. To achieve that, we extend the classic adversarial domain adaptation method to a multiple domain formulation, by using a camera-classifier as discriminator, instead of the usual binary classifier. At the same time, we ensure that our features are ID-discriminative by using a standard classification loss. 34

3.2 Architecture of our camera adversarial transfer strategy. The generator is first updated using source's indentity labels, and minimize the performance of the camera discriminator, using losses (3.8) and (3.9). The discriminator is then updated to recognize from which cameras the feature vectors are extracted with (3.10). 36

3.3 Performance variation with the number of cameras in the setting Market →Duke 46

3.4 Source dataset is in red and target is in blue. Best viewed in color. 47

3.5 The PCA is only done with all target's cameras. 48

3.6 The PCA is only done with target's 7th (purple) and 8th (blue) camera. Best viewed in color. 48
3.7 The PCA is only done with 1st and 2nd target's cameras. 49
3.8 The PCA is only done with 1st and 3rd target's cameras. 49
3.9 The PCA is only done with 2nd and 3rd target's cameras. 50

4.1 Pipeline of our method: alternatively (1) clustering target's training data set using ϕ representation, producing noisy pseudo-label ID $\tilde{p}_n$ alongside centroids ϕ_p, and (2) conditional adversarial training, using a Camera-Discriminator D_{CAM} conditioned by ϕ_p to enforce camera invariance on a per identity basis to avoid negative transfer. Pseudo-label ID are used to train an ID classifier C_{PS-ID} alongside the discriminator. 54
4.2 Mutual information between pseudo labels and camera index evolution for the MMT setting. Ground-truth ID comparison is displayed in dashed lines for both datasets. 63
4.3 Mutual information between pseudo labels and camera index evolution for the SSG [32] setting. Ground-truth ID comparison is displayed in dashed lines for both datasets. 64
4.4 Evolution of the number of lost person IDs during training using MMT on Duke ▶ Mkt. 65
4.5 Evolution of the number of lost person IDs during training using SSG [32] on Duke ▶ Mkt. In SSG, DBSCAN clusters full- (a), upper- (b) and lower-body (c) features independently. 65
4.6 PCA visualization of the embedding for Mkt ▶ Duke setting. Best viewed in color. 68

5.1 Overview of our proposed *Tracking* and *Finetuning* framework: the first step generates pseudo-tracks $\tilde{p}_{s,t}$ on all target sequences s using pretrained tracker weights. The second step finetunes the tracker using source ground truth annotation $p_{s,t}$ and pseudo-tracks $\tilde{b}_{s,t}$ to perform i) Adversarial Sequence Alignment to adapt the detection branch, ii) finetune the ID branch, and iii) enforce Adversarial Detection & ID Disentanglement: i) leverages annotations to extract detection and image features to perform adversarial alignment using D_{image} and D_{ins}, ii) uses a standard classification loss to adapt the ID classifier, and iii) uses a discriminator trained $D_{dis/ID}$ to extract ID information from detection instances and improve disentanglement. GRL are not displayed and considered as first layer of D_*. 72

LIST OF TABLES

2.1 Results on the MOT16 test set, and standard evaluation setting. 28

2.2 Results on MOT17's *moving surveillance camera* setting 29

2.3 Results on MOT17's *robot navigating in the crowd* settings. 29

3.1 R1 and mAP measures on both datasets. The first part of the table is extracted from the literature. All the results are obtained under the same experimental protocol. † refers to unsupervised settings exploiting extra annotations (attributes). Overall best results are shown in **bold**, second best results are in ***italic***. 42

3.2 Adversarial strategies variants. FAT uses only source data at training with $\mathcal{L}_{\text{cam-conf}}$, DAT's discriminator is supervised only with dataset's labels, and *adv* refers to a standard adversarial loss. 44

3.3 Feature's robustness evaluation: we extract features from the gallery images of a given dataset, split them into a testing and training set, and train a camera classifier with the latter on top of the frozen evaluated model. We report the camera accuracy performance on the testing set of each re-ID dataset. Better camera robustness results are in **bold**. Refer to the text for detail. 44

3.4 CAT + LSRO R1 and mAP performance for different number of neurons for the discriminator and μ. We note acc the accuracy performance of the camera discriminator at the end of the training. 45

4.1 Comparison of the proposed **CANU** methodology on the Mkt ► Duke and Duke ► Mkt unsupervised person re-ID settings. **CANU-MMT** establishes a new state-of-the-art in both settings, and **CANU-SGG** outperforms **SSG**. 60

4.2 Comparison of the proposed **CANU** methodology on the Mkt ▶ MSMT and Duke ▶ MSMT unsupervised person re-ID settings. **CANU-MMT** establishes a new state-of-the-art in both settings, and **CANU-SGG** outperforms **SSG**. 61

4.3 Impact of μ in the performance of **CANU**. When the mAP values are equal, we highlight the one corresponding to higher R1. 62

4.4 Evaluation of the impact of the conditional strategy on SGG [32] and MMT [35] (using DSCAN). When the mAP values are equal, we highlight the one corresponding to higher R1. 63

5.1 Domain Adaptation MOT test results for MOT17 $\rightarrow$ MOT20 setting: the official test set of the MOT20 dataset is used. Direct Transfer refers to MOT performance of FairMOT trained on the full training set MOT17. Best results are in **bold**. 78

5.2 Domain Adaptation MOT results for MOT17 $\rightarrow$ MOT20 setting: the second half of of the MOT20 dataset is used as validation set. Direct Transfer refers to MOT performance of FairMOT trained on MOT17. $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$, replacing Adversarial Sequence Alignement, refers to Adversarial Domain Alignment as in [20]. Best results are in **bold**. 79

5.3 Domain Adaptation MOT results for MOT20 $\rightarrow$ MOT17 setting: the second half of of the MOT17 dataset is used as validation set. Direct Transfer refers to MOT performance of FairMOT trained on MOT20. $\mathcal{L}_{dom}^{\mathcal{S}+\mathcal{T}}$, replacing Adversarial Sequence Alignement, refers to Adversarial Domain Alignment as in [20]. Best results are in **bold**. 79

5.4 Domain Adaptation MOT results for MOT20 $\rightarrow$ MOT17 and MOT17 $\rightarrow$ MOT20 settings: the second half of of the target dataset is used as validation set. We change the number of iterations of the *tracking* and *fine-tuning*, ie the number of time η the pseudo-tracks are regenerated during adaptation. The overall number of epochs remains fixed to 20 for all runs. Best results are in **bold**. 80

LIST OF ALGORITHMS

1 Overall tracking algorithm. Updates are performed with the various equations and strategies already described. The frame update counter t_u allows to update the $\phi_{\boldsymbol{w}}$ every T frames. The algorithm outputs the position of all tracks at every frame t. 25

BIBLIOGRAPHY

[1] High quality face recognition with deep metric learning. http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html.

[2] Xavier Alameda-Pineda, Soraya Arias, Yutong Ban, Guillaume Delorme, Laurent Girin, Radu Horaud, Xiaofei Li, Bastien Morgue, and Guillaume Sarrazin. Audio-visual variational fusion for multi-person tracking with robots. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1059–1061, New York, NY, USA, 2019. Association for Computing Machinery.

[3] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *CoRR*, abs/1701.04862, 2017.

[4] Sileye Ba, Xavier Alameda-Pineda, Alessio Xompero, and Radu Horaud. An online variational bayesian model for multi-person tracking from cluttered scenes. *Computer Vision and Image Understanding*, 153:64–76, 2016.

[5] S. Bae and K. Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[6] Nathanael L Baisa. Online multi-object visual tracking using a gm-phd filter with deep appearance learning. In *2019 22th international conference on information fusion (FUSION)*, pages 1–8. IEEE, 2019.

[7] Yutong Ban, Xavier Alameda-Pineda, Fabien Badeig, Sileye Ba, and Radu Horaud. Tracking a varying number of people with a visually-controlled robotic head. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4144–4151. IEEE, 2017.

[8] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Variational bayesian inference for audio-visual tracking of multiple speakers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1761–1776, 2021.

[9] Yutong Ban, Sileye Ba, Xavier Alameda-Pineda, and Radu Horaud. Tracking multiple persons based on a variational bayesian model. In *European Conference on Computer Vision*, pages 52–67. Springer, 2016.

[10] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008.

[12] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[13] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, 2016.

[14] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *2009 IEEE ICCV*, Sep. 2009.

[15] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Partial transfer learning with selective adversarial networks. In *IEEE CVPR*, 2018.

[16] Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M Hospedales. Disjoint label space transfer learning with common factorised space. In *AAAI*, 2019.

[17] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. pages 8866–8875, 06 2020.

[18] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017.

[19] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE TPAMI*, 40(2), 2018.

[20] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *IEEE International Conference on Computer Vision, ICCV*, 10 2017.

[22] Guillaume Delorme, Yutong Ban, Guillaume Sarrazin, and Xavier Alameda-Pineda. ODANet: Online Deep Appearance Network for Identity-Consistent Multi-Person Tracking. In *ICPR 2021 - 25th International Conference on Pattern Recognition / Workshops*, pages 1–15, Milano / Virtual, Italy, January 2021.

[23] Guillaume Delorme, Yihong Xu, Stephane Lathuiliere, Radu Horaud, and Xavier Alameda-Pineda. Canu-reid: A conditional adversarial network for unsupervised person re-identification. In *International Conference on Pattern Recognition*, 2021.

[24] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes, 2020.

[25] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *IEEE CVPR*, 2018.

[26] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2070–2079, 2017.

[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 1996.

[28] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *CoRR*, abs/1705.10444, 2017.

[29] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[30] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9), 2010.

[31] Weitao Feng, Zhihao Hu, Wei Wu, Junjie Yan, and Wanli Ouyang. Multi-object tracking with multiple cues and switcher-aware classification. *arXiv*, 2019.

[32] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *IEEE ICCV*, 2019.

[33] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, 2015.

[34] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1), 2016.

[35] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *ICLR*, 2020.

[36] Kevin Swersky Geoffrey Hinton, Nitish Srivastava. Lecture 6a: Overview of mini-batch gradient descent. http://www.cs.toronto.edu/tijmen/csc321/slides/lecture_slides_lec6.pdf.

[37] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 580–587, June 2014.

[38] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[40] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[41] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *null*, pages 1735–1742. IEEE, 2006.

[42] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[44] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint*, 2017.

[45] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, abs/1704.04861, 2017.

[46] Han-Kai Hsu, Wei-Chih Hung, Hung-Yu Tseng, Chun-Han Yao, Yi-Hsuan Tsai, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[47] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[48] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Deep metric learning for visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(11):2056–2068, 2016.

[49] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[50] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5009, 2018.

[51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *PMLR ICML*, 2015.

[52] N. Jiang, J. Liu, C. Sun, Y. Wang, Z. Zhou, and W. Wu. Orientation-guided similarity learning for person re-identification. In *ICPR*, 2018.

[53] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. In *NIPS*, 2009.

[54] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *CoRR*, abs/2006.02609, 2020.

[55] Mehran Khodabandeh, Arash Vahdat, M. Ranjbar, and W. Macready. A robust learning approach to domain adaptive object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 480–490, 2019.

[56] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014.

[57] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015.

[58] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *IEEE CVPR*, 2012.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[60] Laura Leal-Taixe, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[61] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 772–788, Cham, 2018. Springer International Publishing.

[62] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE CVPR*, 2014.

[63] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *IJCAI*, 2017.

[64] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *The European Conference on Computer Vision (ECCV)*, 2018.

[65] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *IEEE ICCV*, 2019.

[66] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *IEEE CVPR Workshops*, 2018.

[67] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE CVPR*, 2015.

[68] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 1991.

[69] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[70] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[71] Chen Long, Ai Haizhou, Zhuang Zijie, and Shang Chong. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *ICME*, 2018.

[72] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *IEEE CVPR*, 2018.

[73] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7948–7956, 2018.

[74] Tetsu Matsukawa and Einoshin Suzuki. Person re-identification using cnn features learned from combination of attributes. In *ICPR*, 2017.

[75] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *CoRR*, abs/2101.02702, 2021.

[76] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, March 2016. arXiv: 1603.00831.

[77] Anton Milan, Hamid Rezatofighi, Anthony Dick, and Ian Reid. Online multi-target tracking using recurrent neural networks. 04 2016.

[78] Li Minxian, Zhu Xiatian, and Gong Shaogang. Unsupervised person re-identification by deep learning tracklet association. In *ECCV 2018*, 2018.

[79] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, 2014.

[80] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[81] Dang-Khoa Nguyen, Wei-Lun Tseng, and Hong-Han Shuai. Domain-adaptive object detection via uncertainty-aware distribution alignment. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2499–2507, New York, NY, USA, 2020. Association for Computing Machinery.

[82] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018.

[83] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE CVPR*, 2016.

[84] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *IEEE CVPR*, 2019.

[85] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *IEEE ICCV*, 2017.

[86] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[87] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[88] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 3047–3055, 2015.

[89] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016.

[90] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[91] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *IEEE International Conference on Computer Vision, ICCV*, 2017.

[92] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. pages 6949–6958, 06 2019.

[93] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 2020.

[94] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.

[95] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *CoRR*, abs/2012.15460, 2020.

[96] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of Research on Machine Learning Applications*. 2009.

[97] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *IEEE CVPR*, 2017.

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[99] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE CVPR*, 2018.

[100] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *IEEE CVPR*, 2019.

[101] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person trasfer gan to bridge domain gap for person re-identification. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2018.

[102] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE CVPR*, 2016.

[103] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. Springer, 2014.

[104] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11721–11730. IEEE, 2020.

[105] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. Transcenter: Transformers with dense queries for multiple-object tracking. *CoRR*, abs/2103.15145, 2021.

[106] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6787–6796, 2020.

[107] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *CVPR*, 2016.

[108] Yuan Yao, Yu Zhang, Xutao Li, and Yunming Ye. Heterogeneous domain adaptation via soft transfer network. In *ACM MM*, 2019.

[109] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient DETR: improving end-to-end object detector with dense prior. *CoRR*, abs/2104.01318, 2021.

[110] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *IEEE ICCV*, 2017.

[111] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[112] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE ICCV*, 2017.

[113] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. *IEEE CVPR*, 2018.

[114] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *IEEE CVPR*, 2016.

[115] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *IEEE CVPR*, 2019.

[116] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[117] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking, 2020.

[118] Cao Zhangjie, You Kaichao, Long Mingsheng, Wang Jianmin, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *IEEE CVPR*, June 2019.

[119] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE CVPR*, 2017.

[120] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE ICCV*, 2015.

[121] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *CoRR*, abs/1610.02984, 2016.

[122] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE TPAMI*, 35(3), 2013.

[123] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE ICCV*, 2017.

[124] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[125] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE CVPR*, 2019.

[126] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE CVPR*, 2018.

[127] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV 2020*.

[128] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.

[129] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[130] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.